EFFICIENT ESTIMATION IN A REGRESSION MODEL WITH MISSING RESPONSES

A Dissertation

by

SCOTT DANIEL CRAWFORD

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Statistics

EFFICIENT ESTIMATION IN A REGRESSION MODEL WITH MISSING RESPONSES

A Dissertation

by

SCOTT DANIEL CRAWFORD

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Ursula U. Mueller-Harknett
Committee Members,	Jeff Hart
	Soumendra N. Lahiri
	Joel Zinn
Head of Department,	Simon J. Sheather

August 2012

Major Subject: Statistics

ABSTRACT

Efficient Estimation in a Regression Model
with Missing Responses. (August 2012)
Scott Daniel Crawford, B.S., Southern Utah University;
M.S., Brigham Young University;
Chair of Advisory Committee: Dr. Ursula U. Mueller-Harknett

This article examines methods to efficiently estimate the mean response in a linear model with an unknown error distribution under the assumption that the responses are missing at random. We show how the asymptotic variance is affected by the estimator of the regression parameter and by the imputation method. To estimate the regression parameter the Ordinary Least Squares method is efficient only if the error distribution happens to be normal. If the errors are not normal, then we propose a One Step Improvement estimator or a Maximum Empirical Likelihood estimator to estimate the parameter efficiently.

In order to investigate the impact that imputation has on estimation of the mean response, we compare the Listwise Deletion method and the Propensity Score method (which do not use imputation at all), and two imputation methods. We show that Listwise Deletion and the Propensity Score method are inefficient. Partial Imputation, where only the missing responses are imputed, is compared to Full Imputation, where both missing and non-missing responses are imputed. Our results show that in general Full Imputation is better than Partial Imputation. However, when the regression parameter is estimated very poorly, then Partial Imputation will outperform Full Imputation. The efficient estimator for the mean response is the Full Imputation estimator that uses an efficient estimator of the parameter.

ACKNOWLEDGMENTS

A great debt of gratitude goes to my advisor, Dr. Ursula U. Mueller-Harknett, for her help, support, enthusiasm, and interest in my work. She gave me time every week to teach, expound ideas, and broaden my understanding of statistics. Without her help I would never have made sense of that Hajek Le-Cam theory. She is an example of what a researcher should be, and one of a few experts in the field of efficiency and missing data models. Great things can be expected from her and her students in the future.

Thanks to Anton Schick for his amazing insights and expertise with the trickiest forms of the convolution theorem. Also thanks to Wolfgang Wefelmeyer for several professional discussions. Thanks to Paul Martin, to Julie Carroll, to Ellen Toby, and especially Micheal Longnecker for help and encouragement when I needed it most. The entire staff, faculty, and student body of Texas A&M are known for their quality. The Texas A&M University Brazos HPC cluster made the simulations possible.

Thanks to my dear wife, Teasha, who has waited patiently for me to finish this journey and supported me with love and comfort every step of the way. Without Teasha none of my accomplishments would matter, but with her I feel as though I am the most blessed man alive. A thanks to God that sent me such a woman and strengthened me through trials and triumphs alike. I am proud of this work because it is the sum of divine help, my efforts, outstanding academic advising, and the support of friends and family.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF THEOREMS, LEMMAS, AND COROLLARIES

Page

Page

$$\frac{MSE(\hat{\vartheta}_{EFF})}{MSE(\hat{\vartheta}_{OLS})} = \frac{1}{2}.$$

When the missing structure is symmetric over symmetric covariates

$$E(\delta X) = E(\delta)E(X).$$

$$E\{\widehat{E(Y)_{PI}}\} \doteq E(Y).$$

$$\begin{aligned} AV_{PI} &= \vartheta^\top\big\{E(XX^\top) - E(X)E(X)^\top\big\}\vartheta + \sigma^2 E(\delta) \\ &\quad +2E(b\delta\epsilon)^\top\big\{E(X) - E(\delta X)\big\} - 2\vartheta^\top E(\delta X)E(\delta X^\top b) \\ &\quad -2\vartheta^\top E(\delta X b^\top)\big\{E(X) - E(\delta X)\big\} + 2\vartheta^\top E(Xb^\top)E(X) \\ &\quad +\big\{E(X) - E(\delta X)\big\}^\top E(bb^\top)\big\{E(X) - E(\delta X)\big\}. \end{aligned}$$

$$E(\widehat{E(Y)_{FI}}) = E(Y).$$

$$\begin{aligned} AV_{FI} &= \vartheta^\top\big\{E(XX^\top) - E(X)E(X)^\top\big\}\vartheta + 2\vartheta^\top E(Xb^\top)E(X) \\ &\quad +2\vartheta^\top E(X)E(X^\top b) + E(X)^\top E(bb^\top)E(X). \end{aligned}$$

$$\begin{aligned} & n^{1/2}(\hat{\vartheta} - \vartheta) \\ = \ & \big\{E(\delta w_\vartheta(X)X^\top)\big\}^{-1} n^{-1/2} \sum_{i=1}^{n} \delta_i w_\vartheta(X_i)(Y_i - \vartheta^\top X_i) + o_p(1). \end{aligned}$$

Page

Page

$$s_2(\epsilon) + \xi(X, Z, \epsilon).$$

where

$$\begin{aligned} s_2(\epsilon) &\in s(\epsilon) \\ \xi(X, Z, \epsilon) &= \Big[\phi(X, Z) - E\{\phi(X, Z)|\delta = 1\}\Big] l(\epsilon) + E\{\phi(X, Z)|\delta = 1\}\frac{\epsilon}{\sigma^2} \\ \phi(X, Z) &= t^\top \dot{r}_\vartheta(X) + g(Z) \end{aligned}$$

$$E\{\phi(X, Z)\} = E[\delta\{s_{2*}(\epsilon) + \xi_*(X, Z, \epsilon)\}\{s_2(\epsilon) + \xi(X, Z, \epsilon)\}].$$

$$v_*(Y, X, Z) = \xi_*(X, Z, \epsilon).$$

$$E\{\phi(X, Z)\} = E\{\delta\xi_*(X, Z, \epsilon)\xi(X, Z, \epsilon)\}.$$

$$E\{\delta\phi_*(X, Z)\} = \sigma^2.$$

$$\xi_*(X, Z, \epsilon) = \phi_*(X, Z)l(\epsilon) - \frac{\sigma^2}{E(\delta)}l(\epsilon) + \frac{\epsilon}{E(\delta)}.$$

$$E\{\delta l(\epsilon)\xi_*(X, Z, \epsilon)\} = 1.$$

$$E\{\delta\epsilon\xi_*(X, Z, \epsilon)\} = \sigma^2.$$

Page

Page

$$
\begin{aligned}
gr_{(\vartheta_*,\gamma_*)} &= r_\vartheta(X) + \gamma(Z) - E\{r_\vartheta(X) + \gamma(Z)\} + \frac{\delta\epsilon}{E(\delta)} - \frac{\delta l(\epsilon)}{E(\delta)\mathbb{I}} + \frac{\delta l(\epsilon)}{E(\delta|Z)\mathbb{I}} \\
&\quad -\frac{\delta l(\epsilon)}{\mathbb{I}}\left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right]\right) \\
&\quad \left(E\{\delta\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta\dot{r}_\vartheta(X)|Z\}E\{\delta\dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right]\right)^{-1} \\
&\quad \left[\dot{r}_\vartheta(X) - \frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right].
\end{aligned}
$$

CHAPTER I

INTRODUCTION

This work examines methods to efficiently estimate the mean response in a semi-parametric model under the assumption that the responses are missing at random. A study by Elliot (2008) illustrates the complexity of such a problem. He investigated the link between specific minority groups (e.g., non-Mexican Hispanic Americans or Chinese Americans) and obesity in children. The response variable, weight of the child, was frequently missing and laws restricting personal information made it impossible to recover the missing data. Because the missing structure was correlated with other covariates (e.g. height of the child and location) in the model the results would have been biased without imputation, i.e. without estimation of the missing values.

Schick (1993) explains how efficient estimators are formed for regression models when no distributional assumptions are made on the covariates. The *Statistical analysis with missing data* book by Little and Rubin (2002) is well known for its explanation on the estimation of regression parameters under the assumption of data missing at random. Müller et al. (2006) propose the method of full imputation, which estimates all the responses, as an improvement over partial imputation, where only the missing responses are imputed. Müller (2009) showed that in order to efficiently estimate any function of the response it is required to estimate the regression parameters efficiently.

We begin by investigating efficient estimation of the regression parameter when the error distribution is unknown. The Ordinary Least Squares method is proven to be efficient when the error distribution happens to be normal. The complete case

---

The journal model is *The American Statistician*.

versions of the One Step Improvement Estimator discussed in Forrester et al. (2003) and the Maximum Empirical Likelihood Estimator discussed in Peng and Schick (2012) are presented as efficient estimators regardless of the error distribution. We use simulations to show the mean square error of these estimators under various distributions.

To estimate the mean response with missing data we compare four common methods: Listwise Deletion, Propensity Score method, Partial Imputation, and Full Imputation. We also derive the asymptotic variances for each method. Simulation results show the MSE of the estimate of the mean response under various error distributions. We show how the MSE is affected by the method of imputation and by the estimator of the regression parameter.

This research illustrates the impact that the imputation method has on estimation in regression models that have missing data. Full imputation is shown to have the least asymptotic variance when the parameter is estimated efficiently. With an inefficient estimate of the parameter we see that full imputation can have more asymptotic variance than partial imputation. When the missing structure is not symmetric about the covariate, listwise deletion methods will be biased. We find some non-regular errors where the OLS estimator for the regression parameter performs better than efficient estimators. The simulations show how each estimator performs for finite sample sizes.

The paper is organized as follows: Chapter II investigates efficient estimation of the regression parameter. Chapter III shows the asymptotic variance for different methods we use to estimate the mean response with missing data. In Chapter IV we study estimation of the mean response. We compare the asymptotic variance of the partially imputed estimator to the fully imputed estimator. In Chapter V we show the asymptotic variances for the imputation methods under various scenarios. Our

conclusions are in Chapter VI.

Appendix A shows results from simulations using additional models not previously shown. Additionally there are tables of the simulated values for the MSE of estimating ϑ and $E(Y)$. Appendix B contains the R code used to solve equations, create graphs, and run the simulations. Appendix C is an extension of the model given in the paper to a more general semi-parametric model. The method to find the canonical gradient is shown, but no estimator has yet been proven to have an influence function that matches the canonical gradient.

CHAPTER II

PARAMETER ESTIMATION IN LINEAR REGRESSION

A. The model

The model form is

$$Y = \vartheta^\top X + \epsilon$$

where ϑ is assumed to be fixed but unknown, the covariates, X, and errors, ϵ, are assumed to have a random but unknown distribution. The distribution of ϵ, $f(\epsilon)$, has a mean of zero and finite variance σ^2. Further assume that X is square-integrable with finite second moments and independent of ϵ. The expected value of ϵ is 0 with variance σ^2. The observed variables are $(X, \delta, \delta Y)$ where δ is 0 if the response, Y, is not observed, and 1 if the response is observed.

The conditional probability is assumed to depend only on the covariate, not the response, meaning

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = E[\delta|X] = \pi(X).$$

This is called the Missing At Random (MAR) assumption.

The model is studied in Müller et al. (2006). The joint probability of dx, dy, and $d\delta$, $P(dx, dy, d\delta)$ is defined by

$$P(dx, dy, d\delta) = G(dx)B_{\pi(X)}(d\delta)\{\delta Q(x, dy) + (1 - \delta)\delta_o(dy)\}$$

where $G(dx)$ is the marginal distribution on dx, $B_{\pi(X)}(d\delta)$ is the Bernoulli distribution with conditional probability $\pi(X) = P(\delta = 1|X)$, Q is the conditional distribution of

dy given X, and $\delta_o(dy)$ is the Dirac measure at 0.

This work is based on the Hajek-Le Cam theory for locally asymptotically normal families. An estimator of some function $k(G, Q, \pi)$, call it $\hat{k}(G, Q, \pi)$, is asymptotically efficient if it is asymptotically linear, meaning

$$n^{1/2}\{\hat{k}(G, Q, \pi\} - k(G, Q, \pi)) = n^{-1/2} \sum_{i=1}^{n} b(X_i, \epsilon_i, \delta_i) + o_p(1), \tag{2.1}$$

and if $b(X_i, \delta_i Y_i, \delta_i) \in L_{2,0}(P)$ is the efficient influence function.

B. Estimating the parameter

In this section it will be shown that the parameter ϑ in the linear model can be found efficiently, and in Subsection 1 it will be shown that when the unknown error distribution is normal the efficient estimator matches the Ordinary Least Squares estimator. A One Step Improvement estimator is proposed in Subsection 2 and a Maximum Empirical Likelihood estimator is proposed in Subsection 3 which are efficient even when the error distribution is not in fact normal. In Section C simulations are used to find the MSE of each estimator for ϑ under various scenarios. The graphs show how the efficient estimators outperform the inefficient ones for various sample sizes.

To find the influence function for an estimator of the parameter in a parametric model we refer to the work of Müller (2009). Define the parametric model as $Y = r_\vartheta(X) + \epsilon$ where $r_\vartheta(X)$ has derivative $\dot{r}_\vartheta(X)$, and the score function of ϵ is $l(\epsilon)$. Define

$$\zeta(\delta_i, X_i, \epsilon_i) = [\dot{r}_\vartheta(X_i) - E\{\dot{r}_\vartheta(X)|\delta = 1\}]\, l(\epsilon_i) + E\{\dot{r}_\vartheta(X)|\delta = 1\}\frac{\epsilon_i}{\sigma^2}. \tag{2.2}$$

The efficient influence function for the parameter ϑ is

$$b(\delta_i, X_i, \epsilon_i) = E\{\delta\zeta(\delta, X, \epsilon)\zeta(\delta, X, \epsilon)^\top\}^{-1}\delta_i\zeta(\delta_i, X_i, \epsilon_i). \tag{2.3}$$

This influence function holds for any parametric model of Y.

Next we will prove two lemmas. The first lemma shows the influence function under the assumption of linear regression. The second lemma shows the influence function under the assumption of linear regression and normally distributed errors.

Lemma II.1 *Using the model defined above, the influence function for the linear regression model where $r_\vartheta(X) = \vartheta^\top X$ is*

$$b(\delta_i, X_i, \epsilon_i) = E\Bigg[\delta \quad \left[\{X - E(X|\delta=1)\}l(\epsilon) + E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right] \left\{[X - E(X|\delta=1)]l(\epsilon) + E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right\}^\top\Bigg]^{-1} \delta_i\left\{[X_i - E(X|\delta=1)]l(\epsilon_i) + E(X|\delta=1)\frac{\epsilon_i}{\sigma^2}\right\}.$$

PROOF: Assuming $r_\vartheta(X) = \vartheta^\top X$ then $\dot{r}_\vartheta(X) = X$, and Equation 2.2 becomes

$$\zeta(\delta_i, X_i, \epsilon_i) = \{X - E[X|\delta=1]\}\, l(\epsilon_i) + E[X|\delta=1]\frac{\epsilon_i}{\sigma^2}.$$

The efficient influence function from Equation 2.3 is

$$b(\delta_i, X_i, \epsilon_i) = E\Bigg[\delta \quad \left\{[X - E(X|\delta=1)]l(\epsilon) + E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right\} \left\{[X - E(X|\delta=1)]l(\epsilon) + E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right\}^\top\Bigg]^{-1} \delta_i\left\{[X_i - E(X|\delta=1)]l(\epsilon_i) + E(X|\delta=1)\frac{\epsilon_i}{\sigma^2}\right\}. \qquad \blacksquare$$

Lemma II.2 *Using the simple linear regression model when the distribution of the errors, ϵ, is normally distributed an efficient estimator for ϑ will have the asymptotic*

form

$$n^{1/2}(\hat{\vartheta}-\vartheta)=n^{-1/2}E[\delta XX^{\top}]^{-1}\sum_{i=1}^{n}\delta_i X_i\epsilon_i+o_p(1).$$

PROOF: When the errors are normally distributed the score function will be $l(\epsilon)=\frac{\epsilon}{\sigma^2}$. So, by Lemma II.1, the influence function will be

$$\begin{aligned}
b(\delta_i,X_i,\epsilon_i) &= E\Bigg[\delta\left\{[X-E(X|\delta=1)]\frac{\epsilon}{\sigma^2}+E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right\} \\
&\quad \left\{[X-E(X|\delta=1)]\frac{\epsilon}{\sigma^2}+E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right\}^{\top}\Bigg]^{-1} \\
&\quad \delta_i\left\{[X_i-E(X|\delta=1)]\frac{\epsilon_i}{\sigma^2}+E(X|\delta=1)\frac{\epsilon_i}{\sigma^2}\right\} \\
&= E[\delta X\frac{\epsilon}{\sigma^2}X^{\top}\frac{\epsilon}{\sigma^2}]^{-1}\delta_i X_i\frac{\epsilon_i}{\sigma^2} \\
&= E[\delta XX^{\top}\epsilon^2]^{-1}\sigma^2\delta_i X_i\epsilon_i \\
&= E[\delta XX^{\top}]^{-1}E[\epsilon^2]^{-1}\sigma^2\delta_i X_i\epsilon_i \\
&= E[\delta XX^{\top}]^{-1}\delta_i X_i\epsilon_i.
\end{aligned}$$

Putting this influence function into Equation 2.1 with $k(G,Q,\pi)=\vartheta$ gives

$$\begin{aligned}
n^{1/2}(\hat{\vartheta}-\vartheta) &= n^{-1/2}\sum_{i=1}^{n}E[\delta XX^{\top}]^{-1}\delta_i X_i\epsilon_i+o_p(1) \\
&= n^{-1/2}E[\delta XX^{\top}]^{-1}\sum_{i=1}^{n}\delta_i X_i\epsilon_i+o_p(1).
\end{aligned}$$

■

1. Ordinary least squares estimator

In this section we derive the OLS estimator for the missing data model, and introduce a theorem that the OLS estimator is efficient if ϵ is normally distributed. The model under consideration is $Y=\vartheta^{\top}X+\epsilon$ where X is i.i.d. with finite variance and independent of ϵ. Also assume δ is i.i.d. and independent of ϵ and that Y is square

integrable, and $\sum_{i=1}^{n} \delta_i X_i X_i^\top$ is invertible. To start we will introduce a term that is $o_p(1)$ which we will need later.

Lemma II.3 *Using the model and notation above*

$$\left[E[\delta X X^\top]^{-1} - n \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\delta_i X_i \epsilon_i - E[\delta X \epsilon]) = o_p(1).$$

PROOF: By the Weak Law of Large Numbers where $\delta_i X_i X_i^\top$ are i.i.d. with finite variance,

$$\frac{1}{n} \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right] \xrightarrow{p} E[\delta X X^\top].$$

Note that $\delta_i X_i X_i^\top = \sum_{j=1}^{n} \delta_i X_{ij}^2 \geq 0$ is positive semidefinite and $\sum \delta_i X_i X_i^\top$ is invertible by assumption, so by Slutsky's theorem

$$n \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} \xrightarrow{a.s.} E[\delta X X^\top]^{-1}$$

which can be written as

$$E[\delta X X^\top]^{-1} - n \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} = o_p(1). \tag{2.4}$$

Using the assumption of MAR where $E[\delta Y|X] = E[\delta|X]E[Y|X]$

$$\begin{aligned} E[\delta X\epsilon] &= E[\delta X(Y - \vartheta^\top X)] \\ &= E[\delta XY] - E[\delta X\vartheta^\top X] \\ &= E[XE(\delta Y|X)] - E[X\vartheta^\top XE(\delta|X)] \\ &= E[XE(\delta|X)E(Y|X)] - E[X\vartheta^\top XE(\delta|X)] \\ &= E[X\vartheta^\top XE(\delta|X)] - E[X\vartheta^\top XE(\delta|X)] \\ &= 0. \end{aligned} \tag{2.5}$$

Because $\delta_i X_i \epsilon_i$ is i.i.d. with expected value of zero, the Central Limit Theorem states the sum has a limiting normal distribution with a convergence rate of $\sqrt{n}$. This implies

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\delta_i X_i\epsilon_i = O_p(1). \tag{2.6}$$

Then combining Equations 2.4, 2.5, and 2.6

$$\begin{aligned} &\left[E[\delta XX^\top]^{-1} - n\left(\sum_{i=1}^{n}\delta_i X_i X_i^\top\right)^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\delta_i X_i\epsilon_i - E[\delta X\epsilon]) \\ = \;& o_p(1)(O_p(1)) \\ = \;& o_p(1). \end{aligned}$$

■

The Ordinary Least Squares estimator for this model, $\hat{\vartheta}$, minimizes the squared difference between Y and $\hat{Y} = \hat{\vartheta}^\top X_i$. The equation to minimize is

$$\sum_{i=1}^{n}\delta_i(Y_i - \hat{Y}_i)^2.$$

The solution is found by setting the derivative to zero

$$0 = \frac{\partial}{\partial \vartheta} \left(\sum_{i=1}^{n} \delta_i (Y_i - \vartheta^\top X_i)^2 \right) \Big|_{\vartheta = \hat{\vartheta}_{OLS}} = -2 \sum_{i=1}^{n} \delta_i (Y_i - \hat{\vartheta}_{OLS}^\top X_i) X_i^\top .$$

This implies

$$\sum_{i=1}^{n} \delta_i Y_i X_i^\top = \hat{\vartheta}_{OLS}^\top \sum_{i=1}^{n} \delta_i X_i X_i^\top .$$

Solving for $\hat{\vartheta}_{OLS}$ and taking the transpose:

$$\hat{\vartheta}_{OLS} = \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} \sum_{i=1}^{n} \delta_i X_i Y_i .$$

Theorem II.4 *Let $Y = \vartheta^\top X + \epsilon$ where X is i.i.d. and independent of ϵ. Assume that Y is square integrable, and that $\sum_{i=1}^{n} \delta_i X_i X_i^\top$ is invertible. If in fact the errors are normally distributed, $\epsilon \sim N(0, \sigma^2)$, then the OLS estimator $\hat{\vartheta}_{OLS}$ is asymptotically efficient for ϑ. From Equation 2.1 this means*

$$\sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) = n^{-1/2} \sum_{i=1}^{n} b(X_i, \delta_i Y_i, \delta_i) + o_p(1).$$

Proof: The definition of the OLS estimator says

$$\begin{aligned}
& \sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) \\
= \; & \sqrt{n} \left(\left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} \sum_{i=1}^{n} \delta_i X_i Y_i - \vartheta \right) \\
= \; & \sqrt{n} \left(\left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} \left[\sum_{i=1}^{n} \delta_i X_i Y_i \right] - \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right]^{-1} \left[\sum_{i=1}^{n} \delta_i X_i X_i^\top \right] \vartheta \right) .
\end{aligned}$$

This can be written as

$$\begin{aligned}
&\sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) \\
= \;& \sqrt{n}\left(\left[\sum_{i=1}^{n}\delta_i X_i X_i^\top\right]^{-1}\left[\sum_{i=1}^{n}\delta_i X_i(Y_i - X_i^\top\vartheta)\right]\right) \\
= \;& \sqrt{n}\left(\left[\sum_{i=1}^{n}\delta X_i X_i^\top\right]^{-1}\left[\sum_{i=1}^{n}\delta_i X_i\epsilon_i\right]\right).
\end{aligned}$$

Now by Lemma II.3

$$\begin{aligned}
\sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) \;=\; & \left[E[\delta X X^\top]^{-1} - n\left(\sum_{i=1}^{n}\delta_i X_i X_i^\top\right)^{-1}\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\delta_i X_i\epsilon_i - E[\delta X\epsilon]) \\
& + \sqrt{n}\left[\left(\sum_{i=1}^{n}\delta X_i X_i^\top\right)^{-1}\sum_{i=1}^{n}\delta_i X_i\epsilon_i\right] + o_p(1).
\end{aligned}$$

This simplifies to

$$\begin{aligned}
\sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) \;=\; & \frac{1}{\sqrt{n}}E(\delta X X^\top)^{-1}\sum_{i=1}^{n}\delta_i X_i\epsilon_i \\
& - \sqrt{n}\left(\sum_{i=1}^{n}\delta_i X_i X_i^\top\right)^{-1}\sum_{i=1}^{n}\delta_i X_i\epsilon_i \\
& - \sqrt{n}\left[E[\delta X X^\top]^{-1} - n\left(\sum_{i=1}^{n}\delta_i X_i X_i^\top\right)^{-1}\right]E[\delta X\epsilon] \\
& + \sqrt{n}\left(\sum_{i=1}^{n}\delta_i X_i X_i^\top\right)^{-1}\sum_{i=1}^{n}\delta_i X_i\epsilon_i + \\
& + o_p(1)
\end{aligned}$$

This implies

$$\begin{aligned}
\sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) \;&=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(E[\delta X X^\top]\right)^{-1}\delta_i X_i\epsilon_i + 0 + o_p(1) \\
&=\; E[\delta X X^\top]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i X_i\epsilon_i + o_p(1). \qquad (2.7)
\end{aligned}$$

Using Equation 2.1

$$\sqrt{n}(\hat{\vartheta}_{OLS} - \vartheta) = n^{-1/2} \sum_{i=1}^{n} b(X_i, \delta_i Y_i, \delta_i) + o_p(1).$$

Therefore by Lemma II.2 $\hat{\vartheta}_{ols}$ is an efficient estimator for ϑ. ∎

a. OLS with double exponential errors

Consider the model $Y = \vartheta^\top X + \epsilon$ where ϵ has an unknown distribution, but is in fact a Double Exponential random variable, given by the pdf for $z \in \mathbb{R}$, and $\lambda > 0$,

$$f(z) = \frac{1}{2\lambda} e^{-|z|/\lambda}.$$

Further assume Y is MAR depending on δ. Assume X is i.i.d. and independent of ϵ. Assume Y is square integrable and that $\sum_{i=1}^{n} \delta_i X_i X_i^\top$ is invertible. I will use the efficient influence function to show that the OLS estimator is not efficient for all distributions.

Lemma II.5 *For the model desribed above where the distribution of ϵ is unknown, but is actually a Double Exponential random variable the influence function*

$$\begin{aligned} b(\delta_i, X_i, \epsilon_i) \;=\; & E\Bigg[\delta \left(X - \frac{1}{2}E[X|\delta = 1]\right)\left(X - \frac{1}{2}E[X|\delta = 1]\right)^\top \\ & + \frac{1}{4}\delta E[X|\delta = 1]E[X|\delta = 1]^\top\Bigg]^{-1} \\ & \left(sign(\epsilon_i)\delta_i \lambda (X_i - E[X|\delta = 1]) + \frac{1}{2}\delta_i \epsilon_i E[X|\delta = 1]\right). \end{aligned}$$

Proof: As defined above let $f(\epsilon)$ be the distribution of ϵ, and let $\dot{f}(\epsilon)$ be the deriva-

tive, then the score function for the Double Exponential distribution is

$$l(\epsilon) = -\frac{\dot{f}(\epsilon)}{f(\epsilon)} = -\frac{-\frac{sign(\epsilon)}{2\lambda^2}e^{-|\epsilon|/\lambda}}{\frac{1}{2\lambda}e^{-|\epsilon|/\lambda}} = sign(\epsilon)\frac{1}{\lambda}.$$

Using $\sigma^2 = 2\lambda^2$ in Lemma II.1 the influence function is

$$\begin{aligned} b(\delta_i, X_i, \epsilon_i) &= E\Bigg[\delta\left([X - E(X|\delta=1)]\, sign(\epsilon)\frac{1}{\lambda} + E[X|\delta=1]\frac{\epsilon}{2\lambda^2}\right) \\ &\quad \left([X - E(X|\delta=1)]\, sign(\epsilon)\frac{1}{\lambda} + E[X|\delta=1]\frac{\epsilon}{2\lambda^2}\right)^\top\Bigg]^{-1} \\ &\quad \delta_i\left([X_i - E(X|\delta=1)]\, sign(\epsilon_i)\frac{1}{\lambda} + E[X|\delta=1]\frac{\epsilon_i}{2\lambda^2}\right). \end{aligned}$$

Using the fact that $sign(\epsilon)^2 = 1$, $E(\epsilon^2) = 2\lambda^2$ and $E\{sign(\epsilon)\epsilon\} = \lambda$

$$\begin{aligned} b(\delta_i, X_i, \epsilon_i) &= E\Bigg[\delta\Bigg\{\frac{1}{\lambda^2}XX^\top - \frac{1}{\lambda^2}XE(X|\delta=1)^\top + \frac{1}{2\lambda^2}XE(X|\delta=1)^\top \\ &\quad -\frac{1}{\lambda^2}E(X|\delta=1)X^\top + \frac{1}{\lambda^2}E(X|\delta=1)E(X|\delta=1)^\top \\ &\quad -\frac{1}{2\lambda^2}E(X|\delta=1)E(X|\delta=1)^\top + \frac{1}{2\lambda^2}E(X|\delta=1)X^\top \\ &\quad -\frac{1}{2\lambda^2}E(X|\delta=1)E(X|\delta=1)^\top + \frac{1}{2\lambda^2}E(X|\delta=1)E(X|\delta=1)^\top\Bigg\}\Bigg]^{-1} \\ &\quad \delta_i\left\{sign(\epsilon_i)\frac{1}{\lambda}X - sign(\epsilon_i)\frac{1}{\lambda}E(X|\delta=1) + \frac{1}{2\lambda^2}\epsilon_i E(X|\delta=1)\right\} \end{aligned}$$

which simplifies to

$$\begin{aligned} b(\delta_i, X_i, \epsilon_i) &= E\Bigg[\delta\Bigg\{XX^\top - \frac{1}{2}XE(X|\delta=1)^\top - \frac{1}{2}E(X|\delta=1)X^\top \\ &\quad +\frac{1}{2}E(X|\delta=1)E(X|\delta=1)^\top\Bigg\}\Bigg]^{-1} \\ &\quad \delta_i\left[sign(\epsilon_i)\lambda\{X - E(X|\delta=1)\} + \frac{1}{2}\epsilon_i E(X|\delta=1)\right]. \end{aligned}$$

The final form is found by factoring,

$$\begin{aligned} = \; & E\Bigg[\delta\Big\{X-\frac{1}{2}E(X|\delta=1)\Big\}\Big\{X-\frac{1}{2}E(X|\delta=1)\Big\}^\top \\ & +\delta\frac{1}{4}E(X|\delta=1)E(X|\delta=1)^\top\Big\}\Bigg]^{-1} \\ & \delta_i\left[sign(\epsilon_i)\lambda\{X-E(X|\delta=1)\}+\frac{1}{2}\epsilon_i E(X|\delta=1)\right]. \end{aligned}$$

■

By the Hajek-Le Cam Theory the OLS estimator will have a random term which introduces more variability than an efficient estimator. This random term is

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(E[\delta XX^\top]^{-1}\delta_i X_i\epsilon_i - E\Bigg[\delta(X-E[X|\delta=1])(X-E[X|\delta=1])^\top \\ & +\frac{1}{2}\delta E[X|\delta=1]E[X|\delta=1]^\top\Bigg]^{-1} \\ & \Big\{sign(\epsilon_i)\delta_i\lambda(X_i-E[X|\delta=1])+\frac{1}{2}\delta_i\epsilon_i E[X|\delta=1]\Big\}\Bigg). \end{aligned} \tag{2.8}$$

By showing that for a special case scenario this component is not zero, we will show the OLS is not always an efficient estimator. Consider conditionally centered X's where $E[X\delta]=0$. This is equivalent to $E[X|\delta=1]=0$, and Equation 2.8 becomes

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(E[\delta XX^\top]^{-1}\delta_i X_i\epsilon_i - E[\delta XX^\top]^{-1}sign(\epsilon_i)\delta_i\lambda X_i\Big) \\ = \; & \frac{1}{\sqrt{n}}E[\delta XX^\top]^{-1}\sum_{i=1}^{n}\Big(\delta_i X_i(\epsilon_i - sign(\epsilon_i)\lambda)\Big) \\ = \; & \frac{1}{\sqrt{n}}E[\delta XX^\top]^{-1}\sum_{i=1}^{n}\Big(\delta_i X_i(\epsilon_i - \lambda 1_{\epsilon_i>0}+\lambda 1_{\epsilon_i<0})\Big). \end{aligned} \tag{2.9}$$

This does not equal zero, and so the OLS estimator is not efficient for this model. The following Lemma compares the MSE for the efficient estimator with the MSE for the OLS estimator. The MSE for for a random square integrable estimator $\hat{\vartheta}$ is

defined as

$$MSE(\hat{\vartheta}) = E[(\hat{\vartheta} - E(\hat{\vartheta}))(\hat{\vartheta} - E(\hat{\vartheta}))^\top].$$

Lemma II.6 *Under the model where the error term is in fact double exponential and where* $E(X\delta) = 0$ *the ratio of the MSE for the OLS estimator versus the efficient estimator is* $MSE(\hat{\vartheta}_{EFF}) = \frac{1}{2}MSE(\hat{\vartheta}_{OLS})$.

PROOF: First to find the MSE of the OLS estimator using Equation 2.7,

$$\begin{aligned}
nMSE(\hat{\vartheta}_{OLS}) &= E[(\hat{\vartheta}_{OLS} - \vartheta)(\hat{\vartheta}_{OLS} - \vartheta)^\top] \\
&\to E\left[\left(E[\delta XX^\top]^{-1}\delta X\epsilon\right)\left(E[\delta XX^\top]^{-1}\delta X\epsilon\right)^\top\right] \\
&= E\left[E[\delta XX^\top]^{-1}\delta X\epsilon\epsilon X^\top\delta E[\delta XX^\top]^{-1}\right] \\
&= E\left[\epsilon^2 E[\delta XX^\top]^{-1}\delta XX^\top E[\delta XX^\top]^{-1}\right] \\
&= \sigma^2 E[\delta XX^\top]^{-1} = 2\lambda^2 E[\delta XX^\top]^{-1}.
\end{aligned}$$

Now to find the MSE for the efficient estimator using the influence function,

$$\begin{aligned}
nMSE(\hat{\vartheta}_{EFF}) &= E[(\hat{\vartheta}_{EFF} - \vartheta)(\hat{\vartheta}_{EFF} - \vartheta)^\top] \\
&\to E\left[\left(E[\delta XX^\top]^{-1}sign(\epsilon)\delta\lambda X\right)\left(E[\delta XX^\top]^{-1}sign(\epsilon)\delta\lambda X\right)^\top\right] \\
&= E\left[E[\delta XX^\top]^{-1}sign(\epsilon)\delta\lambda XX^\top\lambda\delta sign(\epsilon)E[\delta XX^\top]^{-1}\right] \\
&= E\left[\lambda^2 E[\delta XX^\top]^{-1}\delta XX^\top E[\delta XX^\top]^{-1}\right] \\
&= \lambda^2 E[\delta XX^\top]^{-1}.
\end{aligned}$$

Then the ratio for the MSE for each estimator is

$$MSE(\hat{\vartheta}_{OLS})^{-1}MSE(\hat{\vartheta}_{EFF}) = \frac{\lambda^2 E[\delta XX^\top]^{-1}E[\delta XX]}{2\lambda^2} = \frac{1}{2}.$$

■

This shows that asymptotically the efficient estimator will have half the variance of the OLS estimator.

2. One step improvement estimator

One estimator which is asymptotically efficient for linear regression without missing responses is the One Step Improvement estimator (OSI) described in Forrester et al. (2003). This estimator can be modified analogously for the missing data situation as shown below.

This requires an initial estimate of ϑ, call it $\overline{\vartheta}$, which must be $\sqrt{n}$ consistent and discretized. The Ordinary Least Squares estimator, even if the error distribution is not normal is often used in practice as this original estimate. This estimator is then "improved" by using a Newton-Raphson method with a direct estimator of the influence function. Define

$$\mu = E(X|\delta = 1) \qquad \sigma^2 = E(\epsilon^2).$$

We can estimate μ and σ^2 with $\hat{\mu}$ and $\hat{\sigma^2}$ where

$$\begin{aligned}
\hat{\mu} &= \frac{\sum_{i=1}^n \delta_i X_i}{\sum_{i=1}^n \delta_i} \\
\hat{\sigma^2}_{\overline{\vartheta}} &= \frac{\sum_{i=1}^n \delta_i \epsilon_i(\overline{\vartheta})^2}{\sum_{i=1}^n \delta_i} \\
\epsilon(\vartheta) &= Y - \vartheta^\top X \\
\hat{\zeta}_{\overline{\vartheta}}(X, Y, \delta) &= [X - \hat{\mu}]\hat{l}\{\epsilon(\overline{\vartheta})\} + \hat{\mu}\epsilon(\overline{\vartheta})/\sigma^2_{\overline{\vartheta}}.
\end{aligned}$$

Then the One Step Improvement estimator is

$$\vartheta_{OSI} = \overline{\vartheta} + \left\{ \sum_{i=1}^n \delta_i \hat{\zeta}_{\overline{\vartheta}}(X_i, Y_i, \delta_i) \hat{\zeta}_{\overline{\vartheta}}(X_i, Y_i, \delta_i)^\top \right\}^{-1} \sum_{i=1}^n \delta_i \hat{\zeta}_{\overline{\vartheta}}(X_i, Y_i, \delta_i)$$

where

$$\hat{l}_{\vartheta}(\epsilon) = \frac{-\widehat{\dot{f}}_n(\epsilon)}{\hat{f}_n(\epsilon)}$$

for a kernel based estimate of the error density $\hat{f}_n(\epsilon)$. Call the density estimate $\hat{f}'_n(\epsilon)$. The derivative is estimated as shown in Zhi (2012) using the kernel estimate K with bandwidth h by

$$\hat{f}'_n(\epsilon) = \frac{1}{nh^2}\sum_{i=1}^{n} K'\left(\frac{x - x_i}{h}\right).$$

The OSI estimator is efficient because it uses all the model information, including the independence between ϵ and X.

3. Maximum empirical likelihood estimator

An alternative efficient estimator is the Maximum Empirical Likelihood estimator (MELE) which is explained by Peng and Schick (2012). This method maximizes the empirical likelihood $R_n(\vartheta)$ with respect to ϑ. The MELE estimate is shown to be efficient and equivalent to the OSI estimator in the model with complete observations by Owen (1988). In the case of missing responses the estimator uses the subset of responses which were observed, which is the same subset used for the OSI estimator. Thus, following the reasoning used by Koul et al. (2012), the efficiency is preserved in the MELE estimator. This approach is based on estimating the likelihood in $L_{2,0}$.

The empirical likelihood is

$$\begin{aligned} R_n(\vartheta) \quad = \quad & \sup\Bigg\{\prod_{i=1}^{n} n\pi_i : \pi_i \in [0,1], \sum_{i=1}^{n}\pi_i = 1, \sum_{i=1}^{n}\pi_i\delta_i\epsilon_i(\vartheta) = 0, \\ & \sum_{i=1}^{n}\pi_i\delta_i c_{ik}(\vartheta) = 0, k = 1, \cdots, m\Bigg\}. \end{aligned}$$

The constraint $\sum_{i=1}^{n}\pi_i\delta_i\epsilon_i(\vartheta) = \sum_{i=1}^{n}\pi_i\delta_i(Y_i - \vartheta^\top X) = 0$ comes from the assump-

tion that the errors have mean zero. The m constraints involving c_{ik}'s refer to the independence assumption between X and ϵ. The idea is as follows: independence between X and ϵ implies $E\{(X - EX)a(\epsilon)\} = 0$ for *any* function $a \in L_{2,0}(F)$, where F is the distribution of ϵ. If F is continuous, then $F(\epsilon)$ is uniformly distributed on the interval (0,1). An orthonormal basis of $L_{2,0}(F)$ is $\phi_1 \circ F, \phi_2 \circ F, \cdots$ where the ϕ_k denote a basis of $L_2(U)$. Define $\mathbb{F}_\vartheta$ as the residual based empirical distribution function for F. Then the estimated constraints are

$$c_{ik}(\vartheta) = (X_i - \hat{\mu})\phi_k[\mathbb{F}_\vartheta\{\epsilon_i(\vartheta)\}], k = 1, 2, \cdots, m$$

for some integer $m = m_n \to \infty$ as $n \to \infty$. As shown in Owen (2001) this maximization problem will have exactly one solution in the simple linear case with probability tending to one. For more detail and an implementation of the code in R see the function *el.test* in the R package *emplik*. This estimate incorporates all information in the model rendering it efficient for ϑ.

C. Simulation results for estimating the parameter

To simulate these results let $\vartheta = 3$, $E[\epsilon] = 0$, $X \sim$ Uniform$(0, 2)$, with ϵ and X independent. Then $\hat{\vartheta}$ was calculated using each of the following methods:

1. OLS: Ordinary Least Squares. Estimator will be efficient when the unknown error is in fact normally distributed.

2. OSI: One Step Improvement. Estimator is asymptotically efficient.

3. MELE1: Maximum Empirical Likelihood Estimator with one constraint. For small sample sizes one constraint could be sufficient to achieve efficiency.

4. MELE2: Maximum Empirical Likelihood Estimator with two constraints on the basis. The extra constraint handles larger sample sizes.

5. MELE3: Maximum Empirical Likelihood Estimator with three constraints on the basis. The larger the sample size the more constraints needed to achieve efficiency.

For each method of estimation the data is used only where the response is observed. Thus in this section, we will work under the assumption of no missing data, meaning $\delta_i = 1$ for all i.

The Mean Square Error (MSE) was calculated for each simulation using various methods of estimating ϑ. The methods include Ordinary Least Squares (OLS), One Step Improvement (OSI), and Maximum Empirical Likelihood with one, two, or three constraints on the basis (MELE1, MELE2 and MELE3 respectively).

The efficiency discussed earlier was verified for many error distributions. When the errors are normal all the methods were practically indistinguishable, as expected. Two interesting scenarios are shown. Figure 1 has t_2 errors, and it can be seen that the OLS estimator has a larger MSE, while the other methods are clear improvements. The OSI is worse for small sample sizes because of the difficulty of estimating the score function. Figure 2 has gamma errors shifted to have a mean of zero. The OLS again has more variance, and now the effect of extra constraints in the MELE basis can be seen since MELE3 has the smallest MSE.

For other distributions including the normal which are not shown here see Appendix A.

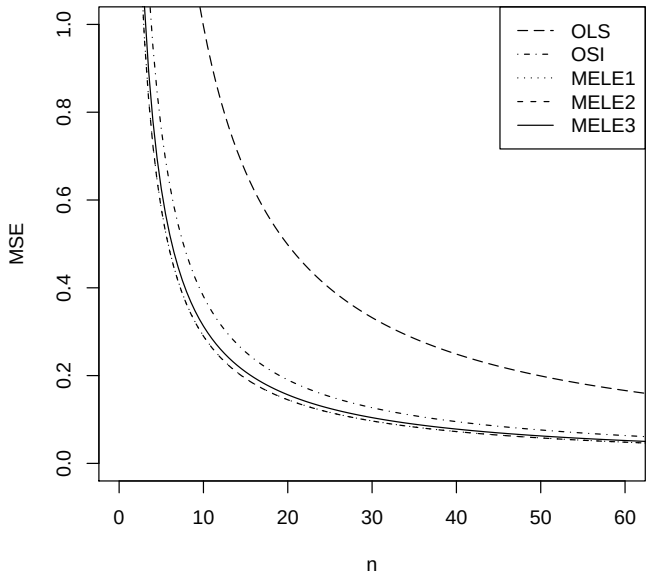

Fig. 1. MSE for various methods of estimating ϑ under t_2 errors.

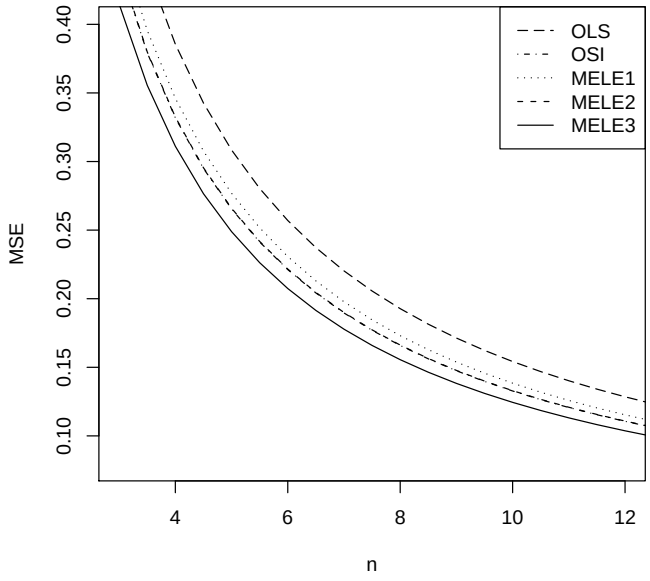

Fig. 2. MSE for various methods of estimating ϑ under gamma errors.

CHAPTER III

ESTIMATING THE MEAN RESPONSE IN REGRESSION

In this section the goal is to estimate $E[Y]$. Define $\widehat{E(Y)}$ as the estimate for an estimator of $E(Y)$. Various estimators will be compared using the asymptotic variance

$$AV = \lim_{n\to\infty} E\left(n\Big[\widehat{E(Y)} - E\Big\{\widehat{E(Y)}\Big\}\Big]^2 \right).$$

The methods that will compared are Listwise Deletion, a Propensity Score method, Partial Imputation, and Full Imputation.

A. Listwise deletion

Listwise Deletion does not use any form of imputation. Any data entry where the response (or the covariate) is missing is simply deleted from the dataset. This is the default method of handling missing data for certain programs. The estimator is

$$\widehat{E(Y)}_{LD} = \frac{\sum_{i=1}^n \delta_i Y_i}{\sum_{i=1}^n \delta_i}. \tag{3.1}$$

This estimator is not consistent. This can be seen from

$$\begin{aligned} E\left\{\widehat{E(Y)}_{LD}\right\} &= E\left(\frac{\sum_{i=1}^n \delta_i Y_i}{\sum_{i=1}^n \delta_i}\right) \\ &= E\left(\frac{\frac{1}{n}\sum_{i=1}^n \delta_i Y_i}{\frac{1}{n}\sum_{i=1}^n \delta_i}\right) \\ &\to E\left\{\frac{E(\delta Y)}{E(\delta)}\right\} \\ &= E(Y|\delta = 1). \end{aligned}$$

The asymptotic result comes from the Law of Large Numbers. This expectation should make sense, since it is the mean of the response where the data is not missing, but this is not in general unbiased for $E[Y]$. There are cases where listwise is unbiased, for example, when the missing structure is symmetric across a symmetric covariate X. Theorem III.1 explains the implication of this assumption.

Theorem III.1 *When the missing structure is symmetric over symmetric covariates*

$$E(\delta X) = E(\delta)E(X).$$

PROOF: For simplicity and reasons of clarity we assume the density of X, namely $g(x)$, is square integrable. Further define the density of X given $\delta = 1$ as $g_1(x)$ and assume it is square integrable. The assumption that the missing structure is symmetric over the symmetric covariate means

$$g_1\{E(X) - x)\} = g_1\{E(X) + x\}. \tag{3.2}$$

Then because $g_1(x)$ is a density,

$$\int_0^\infty g_1\{E(X) + x\}dx + \int_0^\infty g_1\{E(X) - x\}dx = 1,$$

which implies

$$\int_0^\infty g_1\{E(X) + x\}dx = 1/2, \tag{3.3}$$

Because δ is either zero or one, the following holds,

$$\begin{aligned} E(\delta X) &= 0E(X|\delta=0)\{1-E(\delta)\} + 1E(X|\delta=0)E(\delta) \\ &= E(X|\delta=1)E(\delta) \\ &= E(\delta)\int_{-\infty}^{\infty} x g_1(x)dx \\ &= E(\delta)\left\{\int_{-\infty}^{E(X)} x g_1(x)dx + \int_{E(X)}^{\infty} x g_1(x)dx\right\}. \end{aligned}$$

Now changing the indexes so the integrals sum over the same area, but start from $E(X)$ and then x moves toward the limit,

$$\begin{aligned} E(\delta X) &= E(\delta)\left[\int_0^{\infty} \{E(X)-x\} g_1\{E(X)-x\}dx \right. \\ &\qquad \left. + \int_0^{\infty} \{E(X)+x\} g_1\{E(X)+x\}dx\right]. \end{aligned}$$

Using the symmetry of the missing structure from Equation 3.2,

$$\begin{aligned} E(\delta X) &= E(\delta)\left[\int_0^{\infty} \{E(X)-x\} g_1\{E(X)+x\}dx \right. \\ &\qquad \left. + \int_0^{\infty} \{E(X)+x\} g_1\{E(X)+x\}dx\right] \\ &= E(\delta)\left(\int_0^{\infty} \big[\{E(X)-x\} + \{E(X)+x\}\big] g_1\{E(X)+x\}dx\right) \\ &= E(\delta)\int_0^{\infty} 2E(X) g_1\{E(X)+x\}dx \\ &= E(\delta)2E(X)\int_0^{\infty} g_1\{E(X)+x\}dx. \end{aligned}$$

By Equation 3.3

$$\begin{aligned} E(\delta X) &= E(\delta)2E(X)1/2 \\ &= E(\delta)E(X). \end{aligned}$$

■

Using this theorem for Listwise Deletion when the missing structure is symmetric over the covariate then

$$E(X|\delta = 1) = E(X) \iff E(\delta X) = E(1 * X|\delta = 1)E(\delta) = E(\delta)E(X). \tag{3.4}$$

This means for linear regression if the data is missing equally on both sides of $E(X)$ then the estimate of the mean without the missing data will be asymptotically consistent.

The asymptotic variance for Listwise Deletion using Equation 3.1 is

$$\begin{aligned} AV_{LD} &= \lim_{n\to\infty} E\left[n\left\{\frac{\sum_{i=1}^n \delta_i Y_i}{\sum_{i=1}^n \delta_i} - \frac{E(\delta Y)}{E(\delta)}\right\}^2\right] \\ &= \lim_{n\to\infty} nE\left\{\frac{\sum_{i=1}^n \sum_{j=1}^n \delta_i\delta_j Y_i Y_j}{\sum_{i=1}^n \delta_i \sum_{j=1}^n \delta_j} - 2\frac{E(\delta Y)}{E(\delta)}\frac{\sum_{i=1}^n \delta_i Y_i}{\sum_{i=1}^n \delta_i} + \frac{E^2(\delta Y)}{E^2(\delta)}\right\}. \end{aligned}$$

By separating the sums,

$$\begin{aligned} AV_{LD} &= \lim_{n\to\infty} nE\left\{\frac{\sum_{i=1}^n \delta_i Y_i^2 + \sum_{i=1}^n \delta_i Y_i \sum_{j=1,j\neq i}^n \delta_j Y_j}{\sum_{i=1}^n \delta_i + \sum_{i=1}^n \delta_i \sum_{j=1,j\neq i}^n \delta_j}\right. \\ &\quad \left. -2\frac{E(\delta Y)}{E(\delta)}\frac{\sum_{i=1}^n \delta_i Y_i}{\sum_{i=1}^n \delta_i} + \frac{E^2(\delta Y)}{E^2(\delta)}\right\}. \end{aligned}$$

Including the proper constant of n ensures that each summation is an average, which renders the equation in the form

$$\begin{aligned} AV_{LD} &= \lim_{n\to\infty} nE\left\{\frac{\frac{1}{n}\frac{1}{n}\sum_{i=1}^n \delta_i Y_i^2 + \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^n \delta_i Y_i \frac{1}{n-1}\sum_{j=1,j\neq i}^n \delta_j Y_j}{\frac{1}{n}\frac{1}{n}\sum_{i=1}^n \delta_i + \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^n \delta_i \frac{1}{n-1}\sum_{j=1,j\neq i}^n \delta_j}\right. \\ &\quad \left. -2\frac{E(\delta Y)}{E(\delta)}\frac{\frac{1}{n}\sum_{i=1}^n \delta_i Y_i}{\frac{1}{n}\sum_{i=1}^n \delta_i} + \frac{E^2(\delta Y)}{E^2(\delta)}\right\}. \end{aligned}$$

By the law of large numbers each average converges to its finite expected value. By Slutsky's theorem the summations in the equation can be evaluated separately. The

equation then simplifies to

$$\begin{aligned} AV_{LD} &= \lim_{n\to\infty} n\left\{ \frac{\frac{1}{n}E(\delta Y^2) + \frac{n-1}{n}E^2(\delta Y)}{E^2(\delta)} - 2\frac{E^2(\delta Y)}{E^2(\delta)} + \frac{E^2(\delta Y)}{E^2(\delta)} \right\} \\ &= \lim_{n\to\infty} n\left\{ \frac{1}{n}\frac{E(\delta Y^2)}{E^2(\delta)} + \frac{n-1}{n}\frac{E^2(\delta Y)}{E^2(\delta)} - \frac{E^2(\delta Y)}{E^2(\delta)} \right\} \\ &= \frac{E(\delta Y^2)}{E^2(\delta)} - \frac{E^2(\delta Y)}{E^2(\delta)}. \end{aligned}$$

Thus the listwise deletion method is biased unless the missing structure is symmetric across X as shown in Theorem III.1, in which case the asymptotic variance would be

$$AV_{LD} = \frac{\vartheta^\top E(\delta X X^\top)\vartheta}{E^2(\delta)} + \frac{\sigma^2}{E(\delta)} - \vartheta^\top E(X)E(X)^\top\vartheta. \tag{3.5}$$

This variance will be compared with other methods in Chapter V.

B. Propensity score

Another method which does not use imputation but is unbiased is the propensity score method proposed by Rosenbaum and Rubin (1983). For this type of no imputation the data where the response is not observed are deleted from the dataset, but responses which are observed are weighted with the probability of being observed. The estimate for the mean response is

$$\widehat{E(Y)}_{PS} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{E(\delta|X_i)}. \tag{3.6}$$

If $E(\delta|X_i)$ is not known then it can be estimated empirically as done by Dong and Song (2009). It will be assumed that $E(\delta|X)$ is bounded away from zero on the support of X. Note this equation does not require an estimate of ϑ. The estimator

is consistent under the Missing At Random assumption since

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{E(\delta|X_i)}\right\} &= E\left\{\frac{\delta Y}{E(\delta|X)}\right\} \\ &= E\left[E\left\{\frac{\delta Y}{E(\delta|X)}\middle| X\right\}\right] \\ &= E\left\{\frac{E(\delta|X)E(Y|X)}{E(\delta|X)}\right\} \\ &= E(Y). \end{aligned}$$

Deleting the responses which are not observed is a loss of information, and as such the propensity score method is in general not efficient. For a discussion on the disadvantages of such methods see Bell et al. (2009) for details on the bias and variance of such methods.

The asymptotic variance for the propensity score method using Equation 3.6 is

$$\begin{aligned} AV_{PS} &= \lim_{n\to\infty} nE\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{E(\delta|X_i)} - E(Y)\right\}^2\right] \\ &= \lim_{n\to\infty} nE\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j Y_i Y_j}{E(\delta|X_i)E(\delta|X_j)} - 2E(Y)\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{E(\delta|X_i)} + E^2(Y)\right\} \\ &= \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left\{\frac{\delta_i\delta_j Y_i Y_j}{E(\delta|X_i)E(\delta|X_j)}\right\} - 2nE(Y)E\left\{\frac{\delta Y}{E(\delta|X)}\right\} + nE^2(Y)\right]. \end{aligned}$$

By separating the sums

$$\begin{aligned}
AV_{PS} &= \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\delta_i^2 Y_i^2}{E^2(\delta|X_i)}\right\}+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}E\left\{\frac{\delta_i Y_i \delta_j Y_j}{E(\delta|X_i)E(\delta|X_j)}\right\}\right. \\
&\quad \left.-2nE(Y)E\left\{\frac{E(\delta|X)E(Y|X)}{E(\delta|X)}\right\}+nE^2(Y)\right] \\
&= \lim_{n\to\infty}\left[E\left\{\frac{\delta Y^2}{E^2(\delta|X)}\right\}\right. \\
&\quad +(n-1)\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}E\left\{\frac{E(\delta|X_i)E(Y|X_i)E(\delta|X_j)E(Y|X_j)}{E(\delta|X_i)E(\delta|X_j)}\right\} \\
&\quad \left.-2nE^2(Y)+nE^2(Y)\right] \\
&= \lim_{n\to\infty}\left[E\left\{\frac{E(\delta|X)E(Y^2|X)}{E^2(\delta|X)}\right\}+(n-1)E\{E(Y|X_i)E(Y|X_j)\}-nE^2(Y)\right] \\
&= \lim_{n\to\infty}\left[E\left\{\frac{E(Y^2|X)}{E(\delta|X)}\right\}+(n-1)E^2(Y)-nE^2(Y)\right] \\
&= E\left\{\frac{E(Y^2|X)}{E(\delta|X)}\right\}-E^2(Y).
\end{aligned}$$

Now using $Y=\vartheta^\top X+\epsilon$,

$$AV_{PS} = E\left\{\frac{\vartheta^\top XX^\top\vartheta+\sigma^2}{E(\delta|X)}\right\}-\vartheta^\top E(X)E(X)^\top\vartheta. \tag{3.7}$$

This variance will be compared with other methods in Chapter V.

C. Partial imputation

In partial imputation only the missing responses are imputed. The estimate is

$$\widehat{E(Y)_{PI}}=\frac{1}{n}\sum_{i=1}^{n}\Big(\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\Big). \tag{3.8}$$

This method requires an estimate of ϑ. The asymptotic variance of partial imputation depends in part on the variability of $\hat{\vartheta}$. When Y is observed and used instead of

$\hat{\vartheta}X$ then information about the regression structure is lost. This implies partial imputation is not in general efficient.

The efficiency of the estimator used for ϑ will affect the asymptotic variance of the partially imputed estimator. This can be seen by examining the influence function of $\hat{\vartheta}$, $b(\delta, X, \epsilon)$. Assume $\hat{\vartheta}$ is a $\sqrt{n}$ consistent estimator of ϑ. Then the estimate of ϑ can be expanded in asymptotically linear form using Equation 2.1 to

$$\sqrt{n}(\hat{\vartheta} - \vartheta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b(\delta_i, X_i, \epsilon_i) + o_p(1)$$

where $E\{b(\delta_i.X_i, \epsilon_i)\} = 0$.

To simplify the notation the influence function will be written as b_i rather than $b(\delta_i, X_i, \epsilon_i)$. To simplify the equations the notation $\doteq$ will be used to denote equals on the order of $n^{-1/2}$. In other words, $\doteq$ allows suppression of the term $o(n^{-1/2})$. For more detail on the implications of $\sqrt{n}$ estimation in a semiparametric model refer to Schick (1996a), or for a partly linear model refer to Schick (1996b). Then the estimator $\hat{\vartheta}$ can be written as

$$\hat{\vartheta} = \vartheta + \frac{1}{n} \sum_{i=1}^{n} b_i + o_p(n^{-1/2}).$$

Also throughout the following proofs the subscripts X_i vs. X_j will emphasize when two variables are independent of each other. The next theorem shows the partially imputed estimator is asymptotically unbiased.

Theorem III.2 *The expected value for partial imputation is*

$$E\{\widehat{E(Y)_{PI}}\} \doteq E(Y).$$

PROOF: The following expected value will be used frequently:

$$\begin{aligned} E(\hat{\vartheta}^\top X) &\doteq E\left\{\left(\vartheta + \frac{1}{n}\sum_{i=1}^{n} b_i\right)^\top X_j\right\} \\ &\doteq \vartheta^\top E(X) + \frac{n-1}{n} E(b_i^\top X_j) + \frac{1}{n} E(b^\top X) \\ &\doteq \vartheta^\top E(X) + \frac{1}{n} E(b^\top X). \end{aligned} \tag{3.9}$$

By similar proof it can be shown that

$$\begin{aligned} E(\delta\hat{\vartheta}^\top X) &\doteq \vartheta^\top E(\delta X) + \frac{1}{n} E(\delta b^\top X) \\ E(\hat{\vartheta}^\top X X^\top) &\doteq \vartheta^\top E(X X^\top) + \frac{1}{n} E(b^\top X X^\top). \end{aligned}$$

Using these expected values the expected value for partial imputation can be found as

$$\begin{aligned} E\{\widehat{E(Y)_{PI}}\} &= E\left\{\frac{1}{n}\sum_{i=1}^{n} \delta_i Y_i + (1-\delta_i)\hat{\vartheta} X_i^\top\right\} \\ &= E(\delta Y) + E(\hat{\vartheta}^\top X) - E(\delta\hat{\vartheta}^\top X). \end{aligned}$$

Using Equation 3.9,

$$\begin{aligned} E\{\widehat{E(Y)_{PI}}\} &\doteq E\{\delta(\vartheta^\top X + \epsilon)\} + \vartheta^\top E(X) + \frac{1}{n} E(b^\top X) - \vartheta^\top E(\delta X) - \frac{1}{n} E(\delta b^\top X) \\ &\doteq \vartheta^\top E(\delta X) + \vartheta^\top E(X) - \vartheta^\top E(\delta X) + \frac{1}{n} E(b^\top X) - \frac{1}{n} E(\delta b^\top X). \end{aligned}$$

Now using the fact that these results are of order $o_p(n^{-1/2})$,

$$\begin{aligned} E\{\widehat{E(Y)_{PI}}\} &\doteq \vartheta^\top E(X) \\ &\doteq E(\vartheta^\top X + \epsilon) \\ &\doteq E(Y). \end{aligned}$$

■

The next theorem will identify the asymptotic variance.

Theorem III.3 *The asymptotic variance for partial imputation is*

$$\begin{aligned} AV_{PI} &= \vartheta^\top \big\{E(XX^\top) - E(X)E(X)^\top\big\}\vartheta + \sigma^2 E(\delta) \\ &\quad +2E(b\delta\epsilon)^\top \big\{E(X) - E(\delta X)\big\} - 2\vartheta^\top E(\delta X)E(\delta X^\top b) \\ &\quad -2\vartheta^\top E(\delta X b^\top)\big\{E(X) - E(\delta X)\big\} + 2\vartheta^\top E(Xb^\top)E(X) \\ &\quad +\big\{E(X) - E(\delta X)\big\}^\top E(bb^\top)\big\{E(X) - E(\delta X)\big\}. \end{aligned}$$

PROOF: The following expected value will be useful later. The subscripts on X help indicate independent instances of the same random variable.

$$\begin{aligned} E(\hat{\vartheta}^\top X_i X_j^\top) &\doteq E\left\{\left(\vartheta^\top + \frac{1}{n}\sum_{i=1}^{n} b_i^\top\right) X_j X_k^\top\right\} \\ &\doteq E\left(\vartheta^\top X_i X_j^\top + \frac{1}{n} b_i^\top X_i X_j^\top + \frac{1}{n} b_i^\top X_j X_i^\top + \frac{1}{n}\sum_{i\neq \text{i nor j}}^{\text{n-2 terms}} b_i^\top X_j X_k^\top\right) \\ &\doteq \vartheta^\top E(X)E(X)^\top + \frac{1}{n}E(b^\top X)E(X)^\top + \frac{1}{n}E(X)^\top E(bX^\top) \\ &\quad +\frac{n-2}{n}E(b)^\top E(X)E(X)^\top. \end{aligned}$$

Using the fact that $E(b) = 0$,

$$E(\hat{\vartheta}^\top X_i X_j^\top) \doteq \vartheta^\top E(X)E(X)^\top + \frac{1}{n}\left\{E(b^\top X)E(X)^\top + E(X)^\top E(bX^\top)\right\} \quad (3.10)$$

By similar proof

$$\begin{aligned} E(\delta_j \hat{\vartheta}^\top X_i X_j^\top) &\doteq \vartheta^\top E(\delta X)E(X)^\top + \frac{1}{n}\{E(b^\top X)E(\delta X)^\top + E(X)^\top E(\delta bX^\top)\} \\ E(\delta_i\delta_j \hat{\vartheta}^\top X_i X_j^\top) &\doteq \vartheta^\top E(\delta X)E(\delta X)^\top + \frac{1}{n}\left\{E(\delta b^\top X)E(\delta X)^\top + E(\delta X)^\top E(\delta bX^\top)\right\}. \end{aligned}$$

Now an expected value involving $\hat{\vartheta}$ and ϵ,

$$\begin{aligned} E(\epsilon_i \hat{\vartheta}^\top X_j) &\doteq E\left\{ \left(\vartheta^\top + \frac{1}{n} \sum_{i=1}^{n} b_i^\top \right) \epsilon_j X_k \right\} \\ &\doteq E(\epsilon_i \vartheta^\top X_k) + \frac{1}{n} E(b^\top X) E(\epsilon) + \frac{n-2}{n} E(\epsilon) E(b)^\top E(X) + \frac{1}{n} E(b\epsilon)^\top E(X) \\ &\doteq \frac{1}{n} E(b\epsilon)^\top E(X). \end{aligned} \tag{3.11}$$

By a similar proof,

$$\begin{aligned} E(\delta_i \epsilon_i \hat{\vartheta}^\top X_j) &\doteq \frac{1}{n} E(b\delta\epsilon)^\top E(X) \\ E(\delta_i \delta_j \epsilon_i \hat{\vartheta}^\top X_j) &\doteq \frac{1}{n} E(b\delta\epsilon)^\top E(\delta X). \end{aligned}$$

Next is an expected value involving $\hat{\vartheta}^\top X X^\top \hat{\vartheta}$,

$$\begin{aligned} E(\hat{\vartheta}^\top X X \hat{\vartheta}^\top) &\doteq E\left\{ \left(\vartheta^\top + \frac{1}{n} \sum_{i=1}^{n} b_i^\top \right) X_j X_j^\top \left(\vartheta + \frac{1}{n} \sum_{i=1}^{n} b_i \right) \right\} \\ &\doteq E\left(\vartheta^\top X X^\top \vartheta + 2\vartheta^\top \frac{1}{n} \sum_{i=1}^{n} X_j X_j^\top b_i + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} b_i^\top X_k X_k^\top b_j \right) \\ &\doteq \vartheta^\top E(XX^\top)\vartheta + 2\vartheta^\top \frac{1}{n} E(XX^\top b) + 2\vartheta^\top \frac{n-1}{n} E(XX^\top) E(b) \\ &\quad + E\left(\frac{1}{n^2} \overset{\text{n terms}}{\sum_{i=j}} b_i^\top X_k X_k^\top b_i + \frac{1}{n^2} \overset{n^2 - n \text{ terms}}{\sum_{i \neq j}} b_i^\top X_k X_k^\top b_j \right). \end{aligned}$$

Using the fact that $E(b) = 0$,

$$\begin{aligned} E(\hat{\vartheta}^\top XX^\top \hat{\vartheta}) &\doteq \vartheta^\top E(XX^\top)\vartheta + 2\vartheta^\top \frac{1}{n} E(XX^\top b) \\ &\quad + E\Bigg(\frac{1}{n^2} b^\top XX^\top b + \frac{1}{n^2} \overset{\text{n-1 terms}}{\sum_{i=j\neq k}} b_i^\top X_k X_k^\top b_i \\ &\quad + \frac{2(n-1)}{n^2} b_i^\top X_i X_i^\top b_j + \frac{1}{n^2} \overset{(n-1)(n-2) \text{ terms}}{\sum_{i\neq j\neq k}} b_i^\top X_k X_k^\top b_j \Bigg) \\ &\doteq \vartheta^\top E(XX^\top)\vartheta + 2\vartheta^\top \frac{1}{n} E(XX^\top b) \\ &\quad + \frac{1}{n^2} E(b^\top XX^\top b) + \frac{n-1}{n^2} E(b_i^\top X_j X_j^\top b_i) \\ &\quad + \frac{2(n-1)}{n} E(b^\top XX^\top) E(b) + \frac{(n-1)(n-2)}{n^2} E(b)^\top E(XX^\top) E(b). \end{aligned}$$

Again using the $E(b) = 0$ and using the fact that any term of order $1/n^2$ can be included in the term $o(n^{1/2})$ which is noted with the symbol $\doteq$, and denoting the trace of matrix by Tr, then

$$\begin{aligned} E(\hat{\vartheta}^\top XX^\top \hat{\vartheta}) &\doteq \vartheta^\top E(XX^\top)\vartheta + 2\vartheta^\top \frac{1}{n} E(XX^\top b) + \frac{1}{n} E(b_i^\top X_j X_j^\top b_i) \\ &\doteq \vartheta^\top E(XX^\top)\vartheta \\ &\quad + \frac{1}{n}[2\vartheta^\top E(XX^\top b) + Tr\{E(bb^\top) E(XX^\top)\}]. \end{aligned} \tag{3.12}$$

By similar proof

$$E(\delta\hat{\vartheta}^\top XX^\top \hat{\vartheta}) \doteq \vartheta^\top E(\delta XX^\top)\vartheta + \frac{1}{n}[2\vartheta^\top E(\delta XX^\top b) + Tr\{E(bb^\top) E(\delta XX^\top)\}].$$

Another expected value is

$$\begin{aligned} E(\hat{\vartheta}^\top X_i X_j^\top \hat{\vartheta}) &\doteq E\left\{ \left(\vartheta^\top + \frac{1}{n}\sum_{i=1}^{n} b_i^\top \right) X_j X_k^\top \left(\vartheta + \frac{1}{n}\sum_{i=1}^{n} b_i \right) \right\} \\ &\doteq E\Bigg(\vartheta^\top X_i X_j^\top \vartheta + \vartheta^\top \frac{1}{n}\sum_{i=1}^{n} X_j X_k^\top b_i \\ &\quad + \vartheta^\top \frac{1}{n}\sum_{i=1}^{n} X_k X_j^\top b_i + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_i^\top X_k X_L^\top b_j \Bigg). \end{aligned}$$

The sums will be separated into each scenario as defined in the summation,

$$\begin{aligned} E(\hat{\vartheta}^\top X_i X_j^\top \hat{\vartheta}) &\doteq E\Bigg(\vartheta^\top X X^\top \vartheta + \vartheta^\top \frac{2}{n} X_i X_j^\top b_i + \vartheta^\top \frac{2}{n} X_j X_i^\top b_i \\ &\quad + \frac{1}{n^2} \sum_{k\neq i=j\neq L}^{\text{n-2 terms}} b_i^\top X_k X_L^\top b_i + \frac{1}{n^2} \sum_{k=i=j}^{\text{2 terms}} b_i^\top X_i X_k^\top b_i \\ &\quad + \frac{1}{n^2} b_i^\top X_i X_j^\top b_j + \frac{1}{n^2} b_i^\top X_j X_i^\top b_j + \frac{2}{n^2} \sum_{k\neq i\neq j=L}^{\text{n-2 terms}} b_i^\top X_i X_k^\top b_j \\ &\quad + \frac{2}{n^2} \sum_{k\neq i\neq j=L}^{\text{n-2 terms}} b_i^\top X_k X_i^\top b_j + \frac{1}{n^2} \sum_{k\neq i\neq j\neq L}^{n^2-5n+6 \text{ terms}} b_i^\top X_k X_L^\top b_j \Bigg). \end{aligned}$$

Then the expected values are

$$\begin{aligned} E(\hat{\vartheta}^\top X_i X_j^\top \hat{\vartheta}) &\doteq \vartheta^\top E(X) E(X)^\top \vartheta + \vartheta^\top \frac{2}{n} E(Xb^\top) E(X) + \vartheta^\top \frac{2}{n} E(X) E(X^\top b) \\ &\quad + \frac{n-2}{n^2} E(X)^\top E(bb^\top) E(X) + \frac{2}{n^2} E(b^\top X b^\top) E(X) \\ &\quad + \frac{1}{n^2} Tr\{E(bX^\top) E(Xb^\top)\} + \frac{1}{n} Tr\{E(bb^\top) E(X) E(X)^\top\} \\ &\quad + \frac{2n-4}{n^2} E(b^\top X) E(X)^\top E(b) + \frac{2n-4}{n^2} E(X)^\top E(bX^\top) E(b) \\ &\quad + \frac{n^2-5n+6}{n^2} E(b)^\top E(X) E(X)^\top E(b). \end{aligned}$$

Now using $E(b) = 0$ and the fact that any term or order $1/n^2$ is included in the term

$o(n^{-1/2})$ which is denoted by $\doteq$,

$$\begin{aligned} E(\hat{\vartheta}^\top X_i X_j^\top \hat{\vartheta}) &\doteq \vartheta^\top E(X)E(X)^\top \vartheta \\ &\quad + \frac{1}{n}\Big\{ 2\vartheta^\top E(Xb^\top)E(X) + 2\vartheta^\top E(X)E(X^\top b) \\ &\quad + E(X)^\top E(bb^\top)E(X)\Big\}. \end{aligned} \tag{3.13}$$

By a similar proof,

$$\begin{aligned} E(\delta_i \hat{\vartheta}^\top X_i X_j^\top \hat{\vartheta}) &\doteq \vartheta^\top E(X)E(\delta X)^\top \vartheta + \frac{1}{n}\Big\{ \vartheta^\top E(\delta X)E(X^\top b) \\ &\quad + \vartheta^\top E(X)E(\delta X^\top b) + \vartheta^\top E(Xb^\top)E(\delta X) \\ &\quad + \vartheta^\top E(\delta X b^\top)E(X) + E(X)^\top E(bb^\top)E(\delta X)\Big\} \\ E(\delta_i \delta_j \hat{\vartheta}^\top X_i X_j^\top \hat{\vartheta}) &\doteq \vartheta^\top E(\delta X)E(\delta X)^\top \vartheta + \frac{1}{n}\Big\{ 2\vartheta^\top E(\delta X)E(\delta X^\top b) \\ &\quad 2\vartheta^\top E(\delta X b^\top)E(\delta X) + E(\delta X)^\top E(bb^\top)E(\delta X)\Big\}. \end{aligned}$$

The asymptotic variance is defined as

$$\begin{aligned} AV_{PI} &= \lim_{n\to\infty} nE\left(\left[\frac{1}{n}\sum_{i=1}^n \{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\} - E(Y) \right]^2 \right) \\ &= \lim_{n\to\infty} nE\left[\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j + (1-\delta_j)\hat{\vartheta}^\top X_j\} \right] \\ &\quad -n2E(Y)\frac{1}{n}\sum_{i=1}^n \{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\} + nE^2(Y). \end{aligned}$$

The second half of the equation will be simplified first

$$\begin{aligned}
AV_{PI} &= \lim_{n\to\infty}\Bigg(nE\Bigg[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
&\quad -2n\vartheta^\top E(X)E\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}+n\vartheta^\top E(X)E(X)^\top\vartheta\Bigg) \\
&= \lim_{n\to\infty}\Bigg(nE\Bigg[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
&\quad -2n\vartheta^\top E(X)\Big\{E(\delta Y)+E(\hat{\vartheta}^\top X)-E(\delta\hat{\vartheta}^\top X)\Big\} \\
&\quad +n\vartheta^\top E(X)E(X)^\top\vartheta\Bigg).
\end{aligned}$$

Now using Equation 3.9

$$\begin{aligned}
AV_{PI} &= \lim_{n\to\infty}\Bigg(nE\Bigg[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
&\quad -2n\vartheta^\top E(X)\Big\{\vartheta^\top E(\delta X)+\vartheta^\top E(X)+\frac{1}{n}E(b^\top X)-\vartheta^\top E(\delta X) \\
&\quad -\frac{1}{n}E(\delta b^\top X)\Big\}+n\vartheta^\top E(X)E(X)^\top\vartheta\Bigg) \\
&= \lim_{n\to\infty}\Bigg(nE\Bigg[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
&\quad -2n\vartheta^\top E(X)E(\delta X)^\top\vartheta-2n\vartheta^\top E(X)E(X)^\top\vartheta \\
&\quad +2n\vartheta^\top E(X)E(\delta X)^\top\vartheta+n\vartheta^\top E(X)E(X)^\top\vartheta \\
&\quad -2\vartheta^\top E(X)\big\{E(b^\top X)-E(\delta b^\top X)\big\}\Bigg) \\
&= \lim_{n\to\infty}\Bigg(nE\Bigg[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
&\quad -n\vartheta^\top E(X)E(X)^\top\vartheta-2\vartheta^\top E(X)\big\{E(b^\top X)-E(\delta b^\top X)\big\}\Bigg).
\end{aligned}$$

Now the double summation will be simplified by breaking it into two scenarios as

noted by the indicies on the summations,

$$\begin{aligned} AV_{PI} &= \lim_{n\to\infty}\Bigg(nE\Bigg[\frac{1}{n^2}\overset{\text{n terms}}{\sum_{i=j}}\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}^2\Bigg] \\ &\quad + nE\Bigg[\frac{1}{n^2}\overset{n^2-n\text{ terms}}{\sum_{i\neq j}}\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j + (1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\ &\quad - n\vartheta^\top E(X)E(X)^\top\vartheta - 2\vartheta^\top E(X)\big\{E(b^\top X) - E(\delta b^\top X)\big\}\Bigg). \end{aligned} \tag{3.14}$$

The first set of terms in Equation 3.14 where $i = j$ will be simplified separately,

$$\begin{aligned} & E\Bigg[\frac{1}{n^2}\overset{\text{n terms}}{\sum_{i=j}}\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}^2\Bigg] \\ = \;& \frac{n}{n^2}E\Big(\delta Y^2 + 2\delta Y\hat{\vartheta}^\top X - 2\delta Y\hat{\vartheta}^\top X + \hat{\vartheta}^\top XX^\top\hat{\vartheta} - 2\delta\hat{\vartheta}^\top XX^\top\hat{\vartheta} + \delta\hat{\vartheta}^\top XX^\top\hat{\vartheta}\Big) \\ = \;& \frac{1}{n}\Big\{E(\delta Y^2) + 2E(\delta Y\hat{\vartheta}^\top X) - 2E(\delta Y\hat{\vartheta}^\top X) + E(\hat{\vartheta}^\top XX^\top\vartheta) - E(\delta\hat{\vartheta}^\top XX^\top\vartheta)\Big\} \\ = \;& \frac{1}{n}\Big[E\{\delta(\vartheta^\top X + \epsilon)^2\} + E(\hat{\vartheta}^\top XX^\top\vartheta) - E(\delta\hat{\vartheta}^\top XX^\top\vartheta)\Big] \\ = \;& \frac{1}{n}\Big\{\vartheta^\top E(\delta XX^\top)\vartheta + 2\vartheta^\top E(\delta X\epsilon) + E(\delta\epsilon^2) + E(\hat{\vartheta}^\top XX^\top\vartheta) - E(\delta\hat{\vartheta}^\top XX^\top\vartheta)\Big\} \\ = \;& \frac{1}{n}\Big\{\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta + \sigma^2E(\delta) + E(\hat{\vartheta}^\top XX^\top\hat{\vartheta}) - E(\delta\hat{\vartheta}^\top XX^\top\hat{\vartheta})\Big\}. \end{aligned}$$

Using Equation 3.12,

$$\begin{aligned} & E\Bigg[\frac{1}{n}\overset{\text{n terms}}{\sum_{i=j}}\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}^2\Bigg] \\ \doteq \;& \frac{1}{n}\Bigg(\vartheta^\top E(\delta XX^\top)\vartheta + \sigma^2E(\delta) + \vartheta^\top E(XX^\top)\vartheta + \\ & \frac{1}{n}\Big[2\vartheta^\top E(XX^\top b) + Tr\{E(bb^\top)E(XX^\top)\}\Big] \\ & -\vartheta^\top E(\delta XX^\top)\vartheta - \frac{1}{n}\Big[2\vartheta^\top E(\delta XX^\top b) + Tr\{E(bb^\top)E(\delta XX^\top)\}\Big]\Bigg). \end{aligned}$$

Since terms of order $1/n^2$ can be included in the term $o_p(n^{-1/2})$ which is denoted by

the symbol $\doteq$,

$$
\begin{aligned}
& E\left[\frac{1}{n}\sum_{i=j}^{\text{n terms}}\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}^2\right] \\
\doteq\ & \frac{1}{n}\Big\{\vartheta^\top E(\delta X X^\top)\vartheta + \sigma^2 E(\delta) + \vartheta^\top E(XX^\top)\vartheta - \vartheta^\top E(\delta X X^\top)\vartheta\Big\} \\
\doteq\ & \frac{1}{n}\Big\{\sigma^2 E(\delta) + \vartheta^\top E(XX^\top)\vartheta\Big\}. \qquad (3.15)
\end{aligned}
$$

Now the second set of summations used in Equation 3.14 will be simplified,

$$
\begin{aligned}
& E\left[\frac{1}{n^2}\sum_{i\neq j}^{n^2-n\text{ terms}}\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j + (1-\delta_j)\hat{\vartheta}^\top X_j\}\right] \\
=\ & \frac{1}{n^2}E\Bigg\{\sum_{i\neq j}^{n^2-n\text{ terms}}\Big(\delta_i\delta_j Y_i Y_j + 2\delta_i Y_i\hat{\vartheta}^\top X_j - 2\delta_i\delta_j Y_i\hat{\vartheta}^\top X_j + \hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta} \\
& -2\delta_j\hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta} + \delta_i\delta_j\hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta}\Big)\Bigg\} \\
=\ & \frac{n^2-n}{n^2}\Big\{E^2(\delta Y) + 2\vartheta^\top E(\delta_i X_i X_j^\top\hat{\vartheta}) + 2E(\delta_i\epsilon_i\hat{\vartheta}^\top X_j) - 2\vartheta^\top E(\delta_i\delta_j X_i X_j^\top\hat{\vartheta}) \\
& -2E(\delta_i\delta_j\epsilon_i\hat{\vartheta}^\top X_j) + E(\hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta}) - 2E(\delta_j\hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta}) + E(\delta_i\delta_j\hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta})\Big\}
\end{aligned}
$$

Now using the expectations found in Equation 3.10, Equation 3.11, and Equation 3.13,

$$
\begin{aligned}
& E\left[\frac{1}{n^2}\overset{n^2-n \text{ terms}}{\sum_{i\neq j}}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\right] \\
\doteq\ & \frac{n-1}{n}\Bigg(\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta+2\vartheta^\top\Bigg\{E(\delta X)E(X)^\top\vartheta+\frac{1}{n}E(\delta X)E(X^\top b) \\
& +\frac{1}{n}E(\delta X b^\top)E(X)\Bigg\}+2\Bigg\{\frac{1}{n}E(b\delta\epsilon)^\top E(X)\Bigg\}-2\vartheta^\top\Bigg\{E(\delta X)E(\delta X)^\top\vartheta \\
& +\frac{1}{n}E(\delta X)E(\delta X^\top b)+\frac{1}{n}E(\delta X b^\top)E(\delta X)\Bigg\}-2\Bigg\{\frac{1}{n}E(b\delta\epsilon)^\top E(\delta X)\Bigg\} \\
& +\Bigg[\vartheta^\top E(X)E(X)^\top\vartheta+\frac{1}{n}2\vartheta^\top E(X)E(X^\top b)+2\frac{1}{n}\vartheta^\top E(Xb^\top)E(X) \\
& +\frac{1}{n}E(X)^\top E(bb^\top)E(X)\Bigg]-2\Bigg\{\vartheta^\top E(X)E(\delta X)^\top\vartheta+\frac{1}{n}\vartheta^\top E(X)E(\delta X^\top b) \\
& +\frac{1}{n}\vartheta^\top E(\delta X b^\top)E(X)+\frac{1}{n}\vartheta^\top E(\delta X)E(X^\top b)+\frac{1}{n}\vartheta^\top E(Xb^\top)E(\delta X) \\
& +\frac{1}{n}E(X)^\top E(bb^\top)E(\delta X)\Bigg\}+\Bigg\{\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta+\frac{1}{n}2\vartheta^\top E(\delta X)E(\delta X^\top b) \\
& +\frac{1}{n}2\vartheta^\top E(\delta X b^\top)E(\delta X)+\frac{1}{n}E(\delta X)^\top E(bb^\top)E(\delta X)\Bigg\}\Bigg).
\end{aligned}
$$

Since items of order $1/n^2$ can be included in the term $o_p(n^{-1/2})$ which is denoted

using the symbol $\doteq$,

$$
\begin{aligned}
& E\Bigg[\frac{1}{n^2}\sum_{i\neq j}^{n^2-n \text{ terms}}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
\doteq\ & \vartheta^\top E(\delta X)E(\delta X)^\top\vartheta+2\vartheta^\top E(\delta X)E(X)^\top\vartheta-2\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta \\
& +\vartheta^\top E(X)E(X)^\top\vartheta-2\vartheta^\top E(X)E(\delta X)^\top\vartheta+\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta \\
& +\frac{1}{n}\Bigg[-\Big\{\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta+2\vartheta^\top E(\delta X)E(X)^\top\vartheta-2\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta \\
& +\vartheta^\top E(X)E(X)^\top\vartheta-2\vartheta^\top E(X)E(\delta X)^\top\vartheta+\vartheta^\top E(\delta X)E(\delta X)^\top\vartheta\Big\} \\
& +2\vartheta^\top E(\delta X)E(X^\top b)+2\vartheta^\top E(\delta Xb^\top)E(X)+2E(b\delta\epsilon)^\top E(X) \\
& -2\vartheta^\top E(\delta X)E(\delta X^\top b)-2\vartheta^\top E(\delta Xb^\top)E(\delta X)-2E(b\delta\epsilon)^\top E(\delta X) \\
& +2\vartheta^\top E(X)E(X^\top b)+2\vartheta^\top E(Xb^\top)E(X)+E(X)^\top E(bb^\top)E(X) \\
& -2\vartheta^\top E(X)E(\delta X^\top b)-2\vartheta^\top E(\delta Xb^\top)E(X)-2\vartheta^\top E(\delta X)E(X^\top b) \\
& -2\vartheta^\top E(Xb^\top)E(\delta X)-2E(X)^\top E(bb^\top)E(\delta X)+2\vartheta^\top E(\delta X)E(\delta X^\top b) \\
& +2\vartheta^\top E(\delta Xb^\top)E(\delta X)+E(\delta X)^\top E(bb^\top)E(\delta X)\Bigg].
\end{aligned}
$$

This simplifies to

$$
\begin{aligned}
& E\Bigg[\frac{1}{n^2}\sum_{i\neq j}^{n^2-n \text{ terms}}\{\delta_i Y_i+(1-\delta_i)\hat{\vartheta}^\top X_i\}\{\delta_j Y_j+(1-\delta_j)\hat{\vartheta}^\top X_j\}\Bigg] \\
\doteq\ & \vartheta^\top E(X)E(X)^\top\vartheta+\frac{1}{n}\Bigg[-\vartheta^\top E(X)E(X)^\top\vartheta \\
& +2\big\{E(b\delta\epsilon)^\top+\vartheta^\top E(Xb^\top)\big\}\big\{E(X)-E(\delta X)\big\} \\
& +2\vartheta^\top E(X)\big\{E(X^\top b)-E(\delta X^\top b)\big\} \\
& +\big\{E(X)-E(\delta X)\big\}^\top E(bb^\top)\big\{E(X)-E(\delta X)\big\}\Bigg] \qquad (3.16)
\end{aligned}
$$

Now putting Equation 3.16 and Equation 3.15 into Equation 3.14 gives

$$\begin{aligned} AV_{PI} &= \lim_{n\to\infty}\Bigg(\Big\{\sigma^2 E(\delta)+\vartheta^\top E(XX^\top)\vartheta\Big\} \\ &\quad +n\vartheta^\top E(X)E(X)^\top\vartheta+\Big[-\vartheta^\top E(X)E(X)^\top\vartheta \\ &\quad +2\big\{E(b\delta\epsilon)^\top+\vartheta^\top E(Xb^\top)\big\}\big\{E(X)-E(\delta X)\big\} \\ &\quad +2\vartheta^\top E(X)\big\{E(X^\top b)-E(\delta X^\top b)\big\} \\ &\quad +\big\{E(X)-E(\delta X)\big\}^\top E(bb^\top)\big\{E(X)-E(\delta X)\big\}\Big] \\ &\quad -n\vartheta^\top E(X)E(X)^\top\vartheta-2\vartheta^\top E(X)\big\{E(X^\top b)-E(\delta X^\top b)\big\}\Bigg). \end{aligned}$$

This simplifies to

$$\begin{aligned} AV_{PI} &= \vartheta^\top\big\{E(XX^\top)-E(X)E(X)^\top\big\}\vartheta+\sigma^2E(\delta) \\ &\quad +\big\{E(b\delta\epsilon)^\top+\vartheta^\top E(Xb^\top)\big\}\big\{E(X)-E(\delta X)\big\} \\ &\quad +\big\{E(X)-E(\delta X)\big\}^\top E(bb^\top)\big\{E(X)-E(\delta X)\big\}. \end{aligned}$$ ∎

This variance will be compared with full imputation in Chapter IV and compared with other methods in Chapter V.

D. Full imputation

Full imputation incorporates all the information about the model and imputes all the data, even data which are not missing. Full imputation does not imply that data are erased as the observed responses are used in estimating ϑ. As shown in Müller (2009) the formula for full imputation requires weights which account for the error structure. To find the final form of the estimate begin with the equation for full

imputation adjusted for the estimation of $E(Y)$,

$$\widehat{E[Y]_{FI}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} \hat{w}_j \delta_j (\hat{\vartheta}^\top X_i + Y_j - \hat{\vartheta}^\top X_j)}{\sum_{j=1}^{n} \delta_i} \tag{3.17}$$

where $\hat{\epsilon} = Y - \hat{\vartheta}^\top X$, $\hat{w}_j > 0$, and it satisfies the conditions $\sum_{j=1}^{n} \hat{w}_j \delta_j \hat{\epsilon}_j = 0$ and $\sum_{j=1}^{n} \hat{w}_j = n$. As shown by Owen (2001), these weights can be found by

$$\hat{w}_j = \frac{1}{1 + \lambda \delta_j \hat{\epsilon}}.$$

λ is solved by finding the solution to

$$\sum_{j=1}^{n} \frac{\delta_j \hat{\epsilon}_j}{1 + \lambda \delta_j \hat{\epsilon}_j} = 0.$$

Thus when estimating $E(Y)$ the fully imputed estimator given in Equation 3.17 simplifies to

$$\widehat{E[Y]_{FI}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\vartheta}^\top X_i \tag{3.18}$$

The following theorem shows that full imputation is unbiased

Theorem III.4 *The expected value for full imputation is approximately*

$$E(\widehat{E(Y)_{FI}}) \;=\; E(Y).$$

PROOF: Using the expectation found in Equation 3.9,

$$\begin{aligned} E(\widehat{E(Y)_{FI}}) &= E\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\vartheta}^\top X_i\right\} \\ &= E(\hat{\vartheta}^\top X_i) \\ &= \vartheta^\top E(X) + \frac{1}{n}E(b^\top X) \\ &\doteq E(\vartheta^\top X + \epsilon) \\ &\doteq E(Y). \end{aligned}$$

■

Now the asymptotic variance of the fully imputed estimator will be shown.

Theorem III.5 *The asymptotic variance for full imputation is*

$$\begin{aligned} AV_{FI} &= \vartheta^\top\left\{E(XX^\top) - E(X)E(X)^\top\right\}\vartheta + 2\vartheta^\top E(Xb^\top)E(X) \\ &\quad +2\vartheta^\top E(X)E(X^\top b) + E(X)^\top E(bb^\top)E(X). \end{aligned}$$

PROOF: The asymptotic variance is defined as

$$\begin{aligned} AV_{FI} &= \lim_{n\to\infty} nE\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\vartheta}^\top X_i - E(Y)\right\}^2\right] \\ &= \lim_{n\to\infty} nE\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\vartheta}^\top X_i X_j^\top\vartheta - 2E(Y)\frac{1}{n}\sum_{i=1}^{n}\hat{\vartheta}^\top X_i + E^2(Y)\right\} \\ &= \lim_{n\to\infty} nE\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\vartheta}^\top X_i X_j^\top\vartheta - 2E(Y)\left\{E(Y) + \frac{1}{n}E(X^\top b)\right\} + E^2(Y)\right] \\ &= \lim_{n\to\infty} nE\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\vartheta}^\top X_i X_j^\top\vartheta - \vartheta^\top E(X)E(X)^\top\vartheta - 2\frac{1}{n}\vartheta^\top E(X)E(X^\top b)\right\} \end{aligned}$$

The double summations will be split using the same notation as before

$$\begin{aligned} AV_{FI} &= \lim_{n\to\infty} nE\Bigg\{\frac{1}{n^2}\overbrace{\sum_{i=j}}^{\text{n terms}}\hat{\vartheta}^\top X_i X_i^\top \vartheta + \frac{1}{n^2}\overbrace{\sum_{i\neq j}}^{\text{n(n-1) terms}}\hat{\vartheta}^\top X_i X_j^\top \vartheta \\ &\quad -\vartheta^\top E(X)E(X)^\top\vartheta - 2\frac{1}{n}\vartheta^\top E(X)E(X^\top b)\Bigg\}. \end{aligned}$$

Using Equation 3.12 and Equation 3.13,

$$\begin{aligned} AV_{FI} &= \lim_{n\to\infty}\Bigg\{E(\hat{\vartheta}^\top X_i X_i^\top\hat{\vartheta}) + (n-1)E(\hat{\vartheta}^\top X_i X_j^\top\hat{\vartheta}) - n\vartheta^\top E(X)E(X)^\top\vartheta\Bigg\} \\ &= \lim_{n\to\infty}\Bigg\{\Bigg(\vartheta^\top E(XX^\top)\vartheta + \frac{1}{n}\Big[2\vartheta^\top E(XX^\top b) + Tr\{E(bb^\top)E(XX^\top)\}\Big]\Bigg) \\ &\quad +(n-1)\Big[\vartheta^\top E(X)E(X)^\top\vartheta + \frac{1}{n}\{2\vartheta^\top E(Xb^\top)E(X) + 2\vartheta^\top E(X)E(X^\top b) \\ &\quad +E(X)^\top E(bb^\top)E(X)\}\Big] - n\vartheta^\top E(X)E(X)^\top\vartheta - 2\vartheta^\top E(X)E(X^\top b)\Bigg\}. \end{aligned}$$

The terms of order $1/n$ asymptotically approach 0 with probability 1, so

$$\begin{aligned} AV_{FI} &= \vartheta^\top\big\{E(XX^\top) - E(X)E(X)^\top\big\}\vartheta + 2\vartheta^\top E(Xb^\top)E(X) \\ &\quad +E(X)^\top E(bb^\top)E(X). \end{aligned}$$

■

This variance will be compared with partial imputation in Chapter IV. In Chapter V full imputation will be compared with the other estimators discussed earlier. It will be seen that full imputation has the least amount of asymptotic variance when an efficient estimator of ϑ is used.

CHAPTER IV

COMPARISON OF PARTIAL AND FULL IMPUTATION

An estimate is considered efficient if the asymptotically linear form of the estimator matches the efficient influence function as defined in Equation 2.1. Müller (2009) shows that full imputation with an efficient (regular) estimator of ϑ is an ideal method as it is guaranteed to be efficient. The same article in shows in Equation 4.10 what the efficient influence function would be for the parametric model when estimating a generic function $h(X, Y)$. In this paper we have a linear model so the efficient influence function for estimating $E\{h(X, Y)\} = E(Y)$ can be written as

$$\begin{aligned} & n^{1/2}\{\widehat{E(Y)} - E(Y)\} \\ = \; & n^{-1/2}\sum_{i=1}^{n}\Big[\vartheta^\top X_i - \vartheta^\top E(X) \\ & + E(X)^\top E\{\delta\zeta(\delta, X, \epsilon)\zeta(\delta, X, \epsilon)^\top\}^{-1}\delta_i\zeta(\delta_i, X_i, \epsilon_i)\Big] + o_p(1) \end{aligned} \tag{4.1}$$

with $\zeta(\delta, X, \epsilon)$ as defined in Equation 2.2.

Partial Imputation is perhaps more intuitive since only the data that are missing are imputed. The asymptotic variance of each method is determined by the influence function of the estimator used. In this section we explore how the asymptotic variance is affected by an inefficient estimator of ϑ.

A. Efficient estimate for ϑ

As shown in Lemma II.1 when an efficient estimator for ϑ is used the influence function is

$$\begin{aligned} b(\delta_i, X_i, \epsilon_i) = E\Bigg(\delta \quad & \left[\{X - E(X|\delta=1)\}l(\epsilon) + E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right] \\ & \left[\{X - E(X|\delta=1)\}l(\epsilon) + E(X|\delta=1)\frac{\epsilon}{\sigma^2}\right]^\top \Bigg)^{-1} \\ & \delta_i \left[\{X_i - E(X|\delta=1)\}l(\epsilon_i) + E(X|\delta=1)\frac{\epsilon_i}{\sigma^2}\right]. \end{aligned}$$

Using the fact that δ is either 0 or 1, the following simplification can be used:

$$E(X|\delta=1) = \frac{E(1X|\delta=1)E(\delta) + E(0X|\delta=0)\{1-E(\delta)\}}{E(\delta)} = \frac{E(\delta X)}{E(\delta)}.$$

Which means the influence function can be written in a simplified form as

$$b = E(\delta\zeta\zeta^\top)^{-1}\delta\zeta$$

where

$$\zeta = Xl(\epsilon) - \frac{E(\delta X)}{E(\delta)}l(\epsilon) + \frac{\epsilon}{\sigma^2}\frac{E(\delta X)}{E(\delta)}.$$

In this case $E(Xb^\top) = 0$, $E(\delta X b^\top) = 0$, and

$$\begin{aligned} E(bb^\top) &= E\{E(\delta\zeta\zeta^\top)^{-1}\delta\zeta\zeta^\top E(\delta\zeta\zeta^\top)^{-1}\} \\ &= E(\delta\zeta\zeta^\top)^{-1}. \end{aligned}$$

Also

$$\begin{aligned} E(b\delta\epsilon) &= E\left[E(\delta\zeta\zeta^\top)^{-1}\delta\left\{X\epsilon l(\epsilon) - \frac{E(\delta X)}{E(\delta)}\epsilon l(\epsilon) + \frac{\epsilon^2}{\sigma^2}\frac{E(\delta X)}{E(\delta)}\right\}\right] \\ &= E(\delta\zeta\zeta^\top)^{-1}\{E(\delta X) - E(\delta X) + E(\delta X)\} \\ &= E(\delta\zeta\zeta^\top)^{-1}E(\delta X). \end{aligned}$$

It is important to note that $E(\delta\zeta\zeta^\top)$ depends on the fishers information, $\mathbb{I}$, since

$$\begin{aligned} E(\delta\zeta\zeta^\top) &= E\Big\{\delta XX^\top l^2(\epsilon) - \delta X\frac{E(\delta X)^\top}{E(\delta)}l^2(\epsilon) - \delta\frac{E(\delta X)}{E(\delta)}X^\top l^2(\epsilon) \\ &\quad +\delta X\frac{E(\delta X)^\top}{E(\delta)}\frac{1}{\sigma^2}\epsilon l(\epsilon) + \delta\frac{E(\delta X)}{E(\delta)}X^\top\frac{1}{\sigma^2}\epsilon l(\epsilon) + \delta\frac{E(\delta X)E(\delta X)^\top}{E^2(\delta)}l^2(\epsilon) \\ &\quad -\frac{1}{\sigma^2}\delta\frac{E(\delta X)E(\delta X)^\top}{E^2(\delta)}\epsilon l(\epsilon) - \frac{1}{\sigma^2}\delta\frac{E(\delta X)E(\delta X)^\top}{E^2(\delta)}\epsilon l(\epsilon) \\ &\quad +\delta\frac{\epsilon^2}{(\sigma^2)^2}\frac{E(\delta X)E(\delta X)^\top}{E^2(\delta)} \\ &= E(\delta XX^\top)\mathbb{I} - \frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\mathbb{I} - \frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\mathbb{I} \\ &\quad +\frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\frac{1}{\sigma^2} + \frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\frac{1}{\sigma^2} + \frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\mathbb{I} \\ &\quad -\frac{1}{\sigma^2}\frac{E(\delta X)E(\delta X)^\top}{E(\delta)} - \frac{1}{\sigma^2}\frac{E(\delta X)E(\delta X)^\top}{E(\delta)} + \frac{1}{\sigma^2}\frac{E(\delta X)E(\delta X)^\top}{E(\delta)} \\ &= E(\delta XX^\top)\mathbb{I} - \frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\mathbb{I} + \frac{E(\delta X)E(\delta X)^\top}{E(\delta)}\frac{1}{\sigma^2}. \end{aligned} \tag{4.2}$$

Then the asymptotic variance for partial imputation is

$$\begin{aligned} AV_{PI\_EFF} &= \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + \sigma^2E(\delta) \\ &\quad +2E(\delta X)^\top E(\delta\zeta\zeta^\top)^{-1}\{E(X) - E(\delta X)\} \\ &\quad +\{E(X) - E(\delta X)\}^\top E(\delta\zeta\zeta^\top)^{-1}\{E(X) - E(\delta X)\} \\ &= \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + \sigma^2E(\delta) \\ &\quad +E(X)^\top E(\delta\zeta\zeta^\top)^{-1}E(X) - E(\delta X)^\top E(\delta\zeta\zeta^\top)^{-1}E(\delta X). \end{aligned} \tag{4.3}$$

The asymptotic variance for full imputation is

$$AV_{FI\_EFF} = \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + E(X)^\top E(\delta\zeta\zeta^\top)^{-1}E(X) \quad (4.4)$$

It should be noted that this same asymptotic variance can be calculated by taking the expected value of the squared efficient influence function in Equation 4.1.

The difference between the fully imputed estimator and the partially imputed estimator is

$$\begin{aligned} & AV_{PI\_EFF} - AV_{FI\_EFF} \\ = \; & \sigma^2 E(\delta) - E(\delta X)E(\delta\zeta\zeta^\top)^{-1}E(\delta X) \\ = \; & (\sigma^2)^2 E(\delta)\mathbb{I}\Big\{E(\delta) - E(\delta X)^\top E(\delta XX^\top)^{-1}E(\delta X)\Big\} \\ & \Big[\sigma^2\mathbb{I}\{E(\delta) - E(\delta X)^\top E(\delta XX^\top)^{-1}E(\delta X)\} + E(\delta X)^\top E(\delta XX^\top)^{-1}E(\delta X)\Big]^{-1} \end{aligned}$$

To simplify the notation let

$$C_0 = E(\delta) - E(\delta X)^\top E(\delta XX^\top)^{-1}E(\delta X) \quad (4.5)$$

and

$$Q_0 = E(\delta X)^\top E(\delta XX^\top)^{-1}E(\delta X). \quad (4.6)$$

To show that C_0 is nonnegative see the discussion after Equation (2.7) in Müller (2007) to apply the Cauchy Schwarz inequality to higher dimensions. The quantity Q_0 is nonnegative because it has a quadratic form. The difference for the fully imputed estimator and the partially imputed estimator can now be written as

$$\begin{aligned} AV_{PI\_EFF} - AV_{FI\_EFF} &= \sigma^2 E(\delta) - E(\delta X)E(\delta\zeta\zeta^\top)^{-1}E(\delta X) \\ &= (\sigma^2)^2 E(\delta)\mathbb{I}C_0\big(\sigma^2\mathbb{I}C_0 + Q_0\big)^{-1}. \end{aligned}$$

This difference is always nonnegative which implies partial imputation has at least

as much asymptotic variance as full imputation when an efficient estimate for ϑ is used. This supports what the Hajek-Le Cam theory shows in the paper from Müller (2009), which is that full imputation with an efficient estimate for ϑ is efficient.

B. Weighted least squares estimate of ϑ

The model $E(Y|X) = \vartheta^\top X$ suggests estimators for $\hat{\vartheta}$ that solve the equation

$$\sum_{i=1}^{n} \delta_i w_{\hat{\vartheta}}(X_i)(Y_i - \hat{\vartheta}^\top X_i) = 0 \tag{4.7}$$

where $w_{\hat{\vartheta}}$ is a p-dimensional vector of weight functions. Note that $E\{\delta w_\vartheta(X)(Y - \vartheta^\top X)\} = 0$.

The next step is to determine the asymptotic linear form of the weighted least squares estimator. Assumptions IV.1 and IV.2 as well as Theorem IV.3 are based on Section 2 of a paper by Müller (2007).

Assumption IV.1 *The p-dimensional vector $w_\tau(X)$ is $L_2(P)$ differentiable at $\tau = \vartheta$ with a $p \times p$ matrix of partial derivatives $\dot{w}_\vartheta(X)$ and a p-dimensional gradient X, respectively,*

$$E\{|w_\tau(X) - w_\vartheta(X) - \dot{w}_\vartheta(X)(\tau - \vartheta)|^2\} \;=\; o(|\tau - \vartheta|^2).$$

Assumption IV.1 guarantees that the expected value of $w_\tau(X)(Y - \tau^\top X)$ can be approximated as follows,

$$\begin{aligned} & E\{\delta w_\tau(X)(Y - \tau^\top X)\} - E[\delta w_\vartheta(X)\{Y - \vartheta^\top X\}] \\ = \;& -A(\tau - \vartheta) + o(|\tau - \vartheta|), \end{aligned} \tag{4.8}$$

where A is a $p \times p$ matrix of expectations, namely

$$A = E\{\delta w_\vartheta(X) X^\top\} \tag{4.9}$$

Assumption IV.2 *A is invertible.*

By Assumption IV.1, w_τ is $L_2(P)$ differentiable. This implies that the empirical process

$$E_{n\tau} \;=\; n^{-1/2}\sum_{i=1}^{n}\left[\delta_i w_\tau(X_i)(Y_i-\tau^\top X_i) - E\{\delta w_\tau(X)(Y-\tau^\top X)\}\right]$$

is *stochastically equicontinuous* at $\tau=\vartheta$: for every $\varepsilon,\eta>0$ there is a δ such that

$$\limsup_n P(\sup_{|\tau-\vartheta|\le\delta}|E_{n\tau}-E_{n\vartheta}|>\eta)\le\varepsilon. \tag{4.10}$$

See for example Andrews and Pollard (1994) or Müller et al. (2004).

Theorem IV.3 *Any consistent solution $\hat\vartheta$ of Equation 4.7 has the stochastic expansion*

$$\begin{aligned} & n^{1/2}(\hat\vartheta-\vartheta) \\ =\;& \left\{E(\delta w_\vartheta(X)X^\top)\right\}^{-1} n^{-1/2}\sum_{i=1}^{n}\delta_i w_\vartheta(X_i)(Y_i-\vartheta^\top X_i)+o_p(1). \end{aligned} \tag{4.11}$$

Proof: Consider the estimating Equation 4.7 and the empirical process $E_{n\tau}$ from Equation 4.10 in the above remark. We have

$$\begin{aligned} 0 \;&=\; n^{-1/2}\sum_{i=1}^{n}\delta_i w_{\hat\vartheta}(X_i)(Y_i-\hat\vartheta^\top X_i) \\ &=\; E_{n\hat\vartheta}+n^{-1/2}\sum_{i=1}^{n}E\{\delta w_{\hat\vartheta}(X)(Y-\hat\vartheta^\top X)\}+E_{n\vartheta}-E_{n\vartheta} \end{aligned}$$

with $E_{n\hat{\vartheta}} - E_{n\vartheta} = o_p(1)$ by Equation 4.10. Hence

$$\begin{aligned}
0 &= E_{n\vartheta} + n^{-1/2}\sum_{i=1}^{n} E\{\delta w_{\hat{\vartheta}}(X)(Y - \hat{\vartheta}^\top X)\} + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n}\Big[\delta_i w_\vartheta(X_i)(Y_i - \vartheta^\top X_i) \\
&\quad -E\{\delta w_\vartheta(X)(Y - \vartheta^\top X)\} + E\{\delta w_{\hat{\vartheta}}(X)(Y - \hat{\vartheta}^\top X)\}\Big] + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n}\delta_i w_\vartheta(X_i)(Y_i - \vartheta^\top X_i) - An^{1/2}(\hat{\vartheta} - \vartheta) \\
&\quad +n^{1/2}o(|\hat{\vartheta} - \vartheta|) + o_p(1).
\end{aligned}$$

In the last equation we used Equation 4.8. Since the matrix A is invertible by Assumption IV.2 we have proved the desired statement. ∎

Refer to Schick (1996c) for more details on weighted least squares estimation. By Theorem 4.11 the influence function for weighted least squares is

$$b_i = E\{\delta w_\vartheta(X)X^\top\}^{-1}\delta_i w_\vartheta(X_i)(Y_i - \vartheta^\top X_i).$$

For this influence function

$$\begin{aligned}
E(Xb^\top) &= 0 \\
E(\delta Xb^\top) &= 0 \\
E(bb^\top) &= E\Big[E\{\delta w_\vartheta(X)X^\top\}^{-1}\delta w_\vartheta(X)\epsilon\epsilon w_\vartheta(X)^\top E\{\delta X w_\vartheta(X)^\top\}^{-1}\Big] \\
&= \sigma^2 E\{\delta w_\vartheta(X)X^\top\}^{-1}E\{\delta w_\vartheta(X)w_\vartheta(X)^\top\}E\{\delta X w_\vartheta(X)^\top\}^{-1} \\
E(b\delta\epsilon) &= E\Big[E\{\delta w_\vartheta(X)X^\top\}^{-1}\delta w_\vartheta(X)\epsilon^2\Big] \\
&= \sigma^2 E\{\delta w_\vartheta(X)X^\top\}^{-1}E\{\delta w_\vartheta(X)\}.
\end{aligned}$$

By Theorem III.3 the asymptotic variance for partial imputation using weighted least squares is

$$\begin{aligned} AV_{PI\_WLS} &= \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + \sigma^2 E(\delta) \\ &\quad +2\sigma^2 E\{\delta w_\vartheta(X)\}^\top E\{\delta X w_\vartheta(X)^\top\}^{-1}\{E(X) - E(\delta X)\} \\ &\quad +\sigma^2\{E(X) - E(\delta X)\}^\top E\{\delta w_\vartheta(X)X^\top\}^{-1}E\{\delta w_\vartheta(X)w_\vartheta(X)^\top\} \\ &\quad E\{\delta X w_\vartheta(X)^\top\}^{-1}\{E(X) - E(\delta X)\}. \end{aligned} \tag{4.12}$$

ByTheorem III.5 the asymptotic variance for full imputation using weighted least squares is

$$\begin{aligned} AV_{FI\_WLS} &= \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta \\ &\quad +\sigma^2 E(X)^\top E\{\delta w_\vartheta(X)X^\top\}^{-1}E\{\delta w_\vartheta(X)w_\vartheta(X)^\top\} \\ &\quad E\{\delta X w_\vartheta(X)^\top\}^{-1}E(X). \end{aligned} \tag{4.13}$$

The two methods can be compared by determining where the difference between Equation 4.12 and Equation 4.13 is positive or negative. In other words, if

$$\begin{aligned} & AV_{PI\_WLS} - AV_{FI\_WLS} \\ = \; & \sigma^2 E(\delta) + 2\sigma^2 E\{\delta w_\vartheta(X)\}^\top E\{\delta X w_\vartheta(X)^\top\}^{-1}\{E(X) - E(\delta X)\} \\ & -2\sigma^2 E(\delta X)^\top E\{\delta w_\vartheta(X)X^\top\}^{-1}E\{\delta w_\vartheta(X)w_\vartheta(X)^\top\}E\{\delta X w_\vartheta(X)^\top\}^{-1}E(X) \\ & +\sigma^2 E(\delta X)^\top E\{\delta w_\vartheta(X)X^\top\}^{-1} \\ & E\{\delta w_\vartheta(X)w_\vartheta(X)^\top\}E\{\delta X w_\vartheta(X)^\top\}^{-1}E(\delta X) \end{aligned} \tag{4.14}$$

is positive then partial imputation will have a larger asymptotic variance, implying the full imputation is better for that type of weighted least squares estimator. Whether Equation 4.14 is positive or negative depends on the type of weights chosen,

as illustrated by examples in the following sections. Subsection 1 discusses ordinary least squares; other weighted least squares estimators are explored in Subsection 2 and Subsection 3.

1. Ordinary least squares estimate of ϑ

If OLS is used for the estimate for ϑ, then $w_\vartheta(X) = X$, so by Equation 4.12

$$\begin{aligned} AV_{PI\_OLS} &= \vartheta\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + \sigma^2 E(\delta) \\ &\quad + \sigma^2 E(X)^\top E(\delta XX^\top)^{-1} E(X) - \sigma^2 E(\delta X)^\top E(\delta XX^\top)^{-1} E(\delta X) \end{aligned} \quad (4.15)$$

By Equation 4.13 The asymptotic variance for full imputation using OLS with the influence function as shown for the partially imputed estimator is

$$AV_{FI\_OLS} = \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + \sigma^2 E(X)^\top E(\delta XX^\top)^{-1} E(X). \quad (4.16)$$

The difference between the fully imputed estimator and the partially imputed estimator with OLS is

$$AV_{PI\_OLS} - AV_{FI\_OLS} = \sigma^2\{E(\delta) - E(\delta X)^\top E(\delta XX^\top)^{-1} E(\delta X)\}.$$

This is nonnegative by the Cauchy Swarz inequality, so partial imputation will have at least as much asymptotic variance as full imputation using Ordinary Least Squares. In the case where ϵ is normally distributed the OLS estimator will match the efficient estimator. To see this note that under normality $\mathbb{I} = 1/\sigma^2$, which means from Equation 4.2,

$$E(\delta\zeta\zeta^\top) = \frac{1}{\sigma^2} E(\delta XX^\top).$$

Then the asymptotic variance for partial imputation with an efficient estimator comes from Equation 4.3

$$\begin{aligned} AV_{PI\_EFF} &= \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + \sigma^2 E(\delta) \\ &\quad +\sigma^2\{E(X) + E(\delta X)\}^\top E(\delta XX^\top)^{-1}\{E(X) - E(\delta X)\}. \end{aligned}$$

This matches the partial imputation method with an OLS estimator given in Equation 4.15. For full imputation the efficient estimator when $\mathbb{I} = 1/\sigma^2$ is

$$AV_{FI\_EFF} = \vartheta^\top\{E(XX^\top) - E(X)E(X)^\top\}\vartheta + E(\delta X)^\top E(\delta XX^\top)^{-1}E(\delta X).$$

which matches asymptotic variance for the full imputation with an OLS estimator given in Equation 4.16. This shows the approach using the OLS method is asymptotically equivalent to the approach using the efficient method. Consider the case when the error distribution is unknown. We will compare the asymptotic variance of the fully imputed estimator based on the OLS and the fully imputed estimator that uses an efficient estimator. By Equations 4.16 and 4.4 the difference of the asymptotic variances is

$$\begin{aligned} & AV_{FI\_OLS} - AV_{FI\_EFF} \\ = \; & \big\{\sigma^2 E(X)^\top E(\delta XX^\top)^{-1}E(X)\big\} - \big\{E(X)^\top E(\delta\zeta\zeta^\top)^{-1}E(X)\big\}. \end{aligned}$$

Using the notation for Q_0 given in Equation 4.6 and for Cs given in Equation 4.5 this can be written as

$$\begin{aligned} & AV_{FI\_OLS} - AV_{FI\_EFF} \\ = \; & (\sigma^2\mathbb{I} - 1)\sigma^2 Q_0 C_0\big(\sigma^2\mathbb{I}C_0 + Q_0\big)^{-1}. \end{aligned}$$

By the Cauchy Schwarz inequality C_0 is positive, as is the quadratic form Q_0, so this difference is nonnegative when $\mathbb{I} \geq 1/\sigma^2$. This shows that the fully imputed estimator using OLS has in general larger asymptotic variance than the corresponding estimator with an efficient estimator of ϑ. This holds if the usual regularity conditions are met. This will not hold for the uniform distribution, for example, where the support for ϵ depends on the parameter.

2. Constant weight for WLS

If a constant weight, say $w_\vartheta(X) = 1$ is used, then the difference between partial and full imputation given in Equation 4.14 becomes

$$\begin{aligned} & AV_{PI\_WLS} - AV_{FI\_WLS} \\ = \; & \sigma^2 E(\delta) + 2\sigma^2 E(\delta) E(\delta X)^{-1} E(X) - 2\sigma^2 E(\delta) - 2\sigma^2 E(\delta) E(\delta X)^{-1} E(X) + \sigma^2 E(\delta) \\ = \; & 0. \end{aligned}$$

This implies that full imputation and partial imputation are asymptotically equivalent if the weights are constant. This can be seen by rewriting the weighted estimating equation in Equation 4.7,

$$\begin{aligned} \sum_{i=1}^{n} \delta_i (Y_i - \hat{\vartheta}_{w=1}^{\top} X_i) &= 0 \\ \sum_{i=1}^{n} \delta_i Y_i &= \sum_{i=1}^{n} \delta_i \hat{\vartheta}_{w=1}^{\top} X_i \\ \hat{\vartheta}_{w=1} &= \sum_{i=1}^{n} (\delta_i Y_i) \sum_{i=1}^{n} (\delta_i X_i^{\top}) \left\{ \sum_{i=1}^{n} (\delta_i X_i) \sum_{i=1}^{n} (\delta_i X_i^{\top}) \right\}^{-1}. \end{aligned} \tag{4.17}$$

In one dimension this is easier to see the implication of using this weight where $\hat{\vartheta}$ is

$$\hat{\vartheta}_{w=1} = \frac{\sum_{i=1}^{n} \delta_i Y_i}{\sum_{i=1}^{n} \delta_i X_i}.$$

Returning to Equation 3.8, which defines partial imputation,

$$\begin{aligned}\widehat{E(Y)_{PI}} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i Y_i + (1-\delta_i)\hat{\vartheta}_{w=1}^{\top} X_i\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i) + \frac{1}{n}\sum_{i=1}^{n}(\hat{\vartheta}_{w=1}^{\top} X_i) - \frac{1}{n}\sum(\delta_i \hat{\vartheta}_{w=1}^{\top} X_i).\end{aligned}$$

Now using Equation 4.17 in the middle,

$$\begin{aligned}\widehat{E(Y)_{PI}} &= \frac{1}{n}\sum_{i=1}^{n}(\delta_i Y_i) + \frac{1}{n}\sum_{i=1}^{n}(\hat{\vartheta}_{w=1}^{\top} X_i) - \frac{1}{n}\sum(\delta_i Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{\vartheta}_{w=1}^{\top} X_i,\end{aligned}$$

which equals the fully imputed estimator in Equation 3.18 with the estimator having weights equal to one. The proof for any constant weight is similar. Partial imputation is the same as full imputation when a constant weight is used.

3. A poor choice of weights in WLS

Full imputation depends heavily on the estimate for ϑ, so if weighted least squares is used with weights that are chosen poorly enough full imputation will have more asymptotic variance than partial imputation. One example of this is

$$w_{\vartheta}(X) = \frac{X - E(X)}{E(\delta|X)}.$$

In this case

$$\begin{aligned} E(\delta w_\vartheta(X)) &= E\left\{\delta \frac{X - E(X)}{E(\delta|X)}\right\} \\ &= E\left[E\left\{\delta \frac{X - E(X)}{E(\delta|X)}|X\right\}\right] \\ &= E\left\{\frac{X - E(X)}{E(\delta|X)} E(\delta|X)\right\} \\ &= E(X) - E(X) \\ &= 0, \end{aligned}$$

as well as

$$\begin{aligned} E(\delta X w_\vartheta(X)^\top) &= E\left\{\frac{XX^\top - XE(X)^\top}{E(\delta|X)} E(\delta|X)\right\} \\ &= E(XX^\top) - E(X)E(X)^\top, \end{aligned}$$

and

$$\begin{aligned} E(\delta w_\vartheta(X) w_\vartheta(X)^\top) &= E\left[\frac{\{X - E(X)\}\{X - E(X)\}^\top}{E(\delta|X)} E(\delta|X)\right] \\ &= E\{XX^\top - 2XE(X)^\top + E(X)E(X)^\top\} \\ &= E(XX^\top) - E(X)E(X)^\top. \end{aligned}$$

The difference in asymptotic variance between partial and full imputation using Equation 4.14 is

$$\begin{aligned} & AV_{PI\_WLS} - AV_{FI\_WLS} \\ = \; & \sigma^2 E(\delta) - 2\sigma^2 E(\delta X)^\top \{E(XX^\top) - E(X)E(X)^\top\}^{-1} E(X) \\ & + \sigma^2 E(\delta X)^\top \{E(XX^\top) - E(X)E(X)^\top\}^{-1} E(\delta X) \\ = \; & \sigma^2 E(\delta) + \sigma^2 E(\delta X)^\top \{E(XX^\top) - E(X)E(X)^\top\}^{-1} \{E(\delta X) - 2E(X)\} \end{aligned} \quad (4.18)$$

When the difference is negative partial imputation has less variance than full impu-

tation. The following explanation shows that this scenario is possible. The method is to increase the value of Equation 4.18, but then show a scenario where the larger value is still negative. That implies Equation 4.18 would be negative and partial imputation is better than full imputation in that case. Consider the inequality

$$E(\delta X)\{Cov(X)\}^{-1}E(\delta X) \leq E(\delta X)\{Cov(X)\}^{-1}E(X). \tag{4.19}$$

When X is positive Equation 4.19 is true. When X is negative then the two expected values in each side of the equation net a postive result, so Equation 4.19 is still true. This means Equation 4.18 can be increased by

$$\begin{aligned} & \sigma^2 E(\delta) + \sigma^2 E(\delta X)^\top \{E(XX^\top) - E(X)E(X)^\top\}^{-1}\{E(\delta X) - 2E(X)\} \\ \leq \;& \sigma^2 E(\delta) - \sigma^2 E(\delta X)^\top \{E(XX^\top) - E(X)E(X)^\top\}^{-1}E(\delta X). \end{aligned}$$

If the larger equation is still negative, then partial imputation has less variability than full imputation. In other words, if

$$0 \;\geq\; E(\delta) - E(\delta X)^\top \{E(XX^\top) - E(X)E(X)^\top\}^{-1}E(\delta X).$$

In a one dimensional case this is means partial imputation has less variability than full imputation if

$$var(X) \leq \frac{E^2(\delta X)}{E(\delta)}. \tag{4.20}$$

To see that this scenario is possible consider the case when the missing structure is symmetric over a symmetric covariate. This assumption as shown in Theorem III.1 implies

$$E(\delta X) = E(\delta)E(X).$$

Then Equation 4.20 becomes

$$E(X^2) \leq \{E(\delta) + 1\}E^2(X).$$

Such a scenario is explored in Section V. In that case the estimate of ϑ has so much variability that full imputation has more asymptotic variance than partial imputation.

CHAPTER V

EXAMPLES

In this section the goal is to estimate $E(Y)$ under the simple linear model where $Y = \vartheta X + \epsilon$. Let $X \sim\ Uniform(0, 2)$, and $\vartheta = 3$. The following error distributions of ϵ are considered:

1. $\epsilon \sim\ Uniform(-1, 1)$ which breaks the regularity conditions;
2. $\epsilon \sim\ Normal(0, 1)$ where the OLS method is efficient;
3. $\epsilon \sim\ Double\ Exponential$ with a mean of 0 and variance 2;
4. $\epsilon \sim\ t$ with 3 degrees of freedom;
5. $\epsilon \sim\ Gamma$ with a variance of 2 shifted to have a mean of 0.
6. $\epsilon \sim\ Logistic$. The standard Logistic distribution has heavy tails compared to the Normal distribution.
7. $\epsilon \sim\ Gumbel$. The standard Gumbel distribution is similar to the Normal but is skewed.

Table I lists some calculations that are helpful for finding the asymptotic variance under each error distribution.

For each scenario the asymptotic variances will be compared using the following methods:

1. LD - Listwise Deletion (Equation 3.5);
2. PS - Propensity Score using the true $\pi(X)$ (Equation 3.7);
3. PI_EFF - Partial Imputation with an efficient estimate (Equation 4.3);

Table I. The variance and Fisher's information for various error distributions.

	σ^2	$\mathbb{I}$
U(-1,1)	.333	0
N(0,1)	1	1
t_3	0.6667	3
Gamma(2,1)-2	30.474	2
Logistic	0.3333	3.2899
Gumbel	1	1.64493
DExp	2	1

4. FI_EFF - Full imputation with an efficient estimate (Equation 4.4);

5. PI_OLS - Partial Imputation using Ordinary Least Squares (Equation 4.15);

6. FI_OLS - Full Imputation using Ordinary Least Squares (Equation 4.16).

A. Symmetric missing structure

Consider the case where the missing structure $\pi(X) = E(\delta|X)$ has the piecewise form

$$\pi(X) = \begin{cases} .2 & 0 < X < .5 \\ .8 & .5 \leq X < 1.5 \\ .2 & 1.5 \leq X \leq 2 \end{cases}$$

which means δ is 1 with probability of 0.20 on the ends of X, and is 1 with probability 0.80 in the middle range of X. Figure 3 shows a plot of $\pi(X)$.

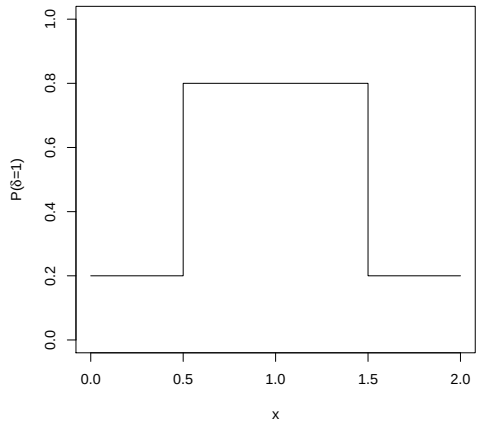

Fig. 3. Symmetric missing structure. The missingness is centered over $E(X) = 1$ and is stepwise.

It is easy to show that $E(X) = 1$, $E(X^2) = 4/3$, and

$$E(\delta) = (0.2)(1/4) + (0.8)(1/2) + (0.2)(1/4) = 1/2.$$

The following expectation is also needed:

$$\begin{aligned} E(\delta X) &= \int_0^1 \int_0^2 \delta x G(dx) B_{\pi(x)}(d\delta) \\ &= (0.2) \int_0^{0.5} x \frac{1}{2} dx + (0.8) \int_{0.5}^{1.5} x \frac{1}{2} dx + (0.2) \int_{1.5}^{2} x \frac{1}{2} dx \\ &= 0.5. \end{aligned}$$

And by using the same method $E(\delta X^\top X) = 71/120 \approx 0.59167$. The next expectation uses the fact that $E(\delta|X) = (0.2)1(0 \leq X \leq 0.5) + (0.8)1(0.5 < X \leq 1.5) + (0.2)1(1.5 < X \leq 2)$. This means

$$\begin{aligned} E\left\{\frac{1}{E(\delta|X)}\right\} &= \int_0^{0.5} \frac{1}{0.2}\frac{1}{2} dx + \int_{0.5}^{1.5} \frac{1}{0.8}\frac{1}{2} dx + \int_{1.5}^{2} \frac{1}{0.2}\frac{1}{2} dx \\ &= \frac{25}{8} = 3.125. \end{aligned}$$

Also,

$$\begin{aligned} E\left\{\frac{X^2}{E(\delta|X)}\right\} &= \int_0^{0.5} \frac{1}{0.2}x^2\frac{1}{2}dx + \int_{0.5}^{1.5} \frac{1}{0.8}x^2\frac{1}{2}dx + \int_{1.5}^{2} \frac{1}{0.2}x^2\frac{1}{2}dx \\ &= \frac{445}{96} \approx 4.635417. \end{aligned}$$

This specific model is interesting because $E(\delta X) = E(\delta)E(X)$. See Theorem III.1 for an explanation of how this implies the listwise deletion method is not biased. The asymptotic variances are given in Table II. It is easy to see that the Propensity Score method which uses the true $\pi(X)$ has the highest variance. The Listwise Deletion method has less variance than the propensity score method because the error structure is symmetric over $E(X)$. When the error structure is uniform the regularity conditions do not hold, and there is no regular efficient estimator for the mean response. In this case the estimates using an OLS estimate for ϑ have the least variance. When the errors are normal the imputation methods to estimate the mean response have the same asymptotic variance whether an OLS estimate or the efficient estimate of ϑ is used. See Subsection 1 for discussion on how the OLS estimate is efficient under normality. In every other case using full imputation with the efficient estimate of ϑ results in the smallest asymptotic variance.

Now consider weighted least squares where the choice of weights follows Subsection 3,

$$w_\vartheta(X) = \frac{X - E(X)}{\pi(X)}. \tag{5.1}$$

The asymptotic variances for partial and full imputation when the errors are standard normally distributed are

$$\begin{aligned} AV_{PI\_WLSbad} &= 997.5 \\ AV_{FI\_WLSbad} &= 998.0 \end{aligned}$$

Table II. The asymptotic variances where the missing structure is symmetric.

	LD	PS	PI_OLS	PI_EFF	FI_OLS	FI_EFF
U(-1,1)	13.0	33.8	3.59	3.67	3.56	3.67
N(0,1)	14.3	35.8	4.77	4.77	4.69	4.69
Dexp	16.3	39.0	6.54	6.20	6.38	5.23
t_3	18.3	42.1	8.30	7.79	8.07	7.39
Gamma(2,1)-2	16.3	39.0	6.54	6.20	6.38	5.93
Logistic	18.9	43.0	8.82	8.75	8.56	8.48
Gumbel	15.6	37.9	5.91	5.72	5.78	5.53

As discussed in Subsection 3, this version of weighted least squares causes full imputation to have more asymptotic variance than partial imputation. From Equation 4.20 it can be seen that this difference would be even more drastic if $\pi(X)$ was closer to zero or if the variance of X was smaller.

B. Gaussian missing structure

Now consider a Gaussian missing structure where

$$\pi(Xa) = E(\delta|X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(X-\mu_X)^2}{\sigma_X}}.$$

This model is similar to the symmetric missing structure discussed earlier, but it is a smooth function. The plot for the Gaussian missing structure is given in Figure 4.

As before $E(X) = 1$, $E(X^2) = 4/3$, and $\sigma_X = 1/3$. The other needed pieces are

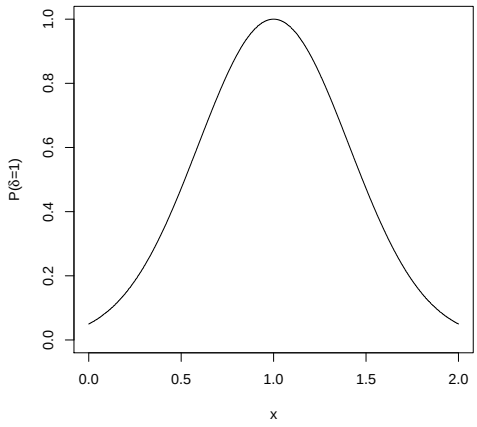

Fig. 4. Gaussian missing structure.

solved as

$$\begin{aligned}
E(\delta) &= \int_0^2 \frac{1}{2} e^{-3(X-1)^2} dX = 0.5043435 \\
E(\delta X) &= \int_0^2 X \frac{1}{2} e^{-3(X-1)^2} dX = 0.5043435 \\
E(\delta X^2) &= \int_0^2 X^2 \frac{1}{2} e^{-3(X-1)^2} dX = 0.5801029 \\
E\left\{\frac{1}{E(\delta|X)}\right\} &= \int_0^2 \frac{1}{2} e^{3(X-1)^2} dX = 4.2222 \\
E\left\{\frac{X^2}{E(\delta|X)}\right\} &= \int_0^2 \frac{1}{2} X^2 e^{3(X-1)^2} dX = 6.8661
\end{aligned}$$

The asymptotic variances for this model are given in Table III. The results are similar to the symmetric model. The uniform distribution shows a smaller asymptotic variance for methods that use the OLS estimator for ϑ, and under the normal distribution the OLS and efficient estimates for ϑ agree. For every other error distribution the fully imputed estimator with an efficient estimator of ϑ has the smallest asymptotic variance.

Table III. The asymptotic variances where the missing structure is Gaussian.

	LD	PS	PI_EFF	FI_EFF	PI_OLS	FI_OLS
U(-1,3)	12.1	54.2	3.661	3.661	3.597	3.575
Normal	13.5	57.0	4.79	4.724	4.79	4.724
DExp	15.5	61.2	6.282	6.049	6.579	6.448
t_3	17.5	65.5	7.924	7.574	8.369	8.171
Gamma(2,1)-2	15.5	61.2	6.282	6.049	6.579	6.448
Logistic	18.0	66.7	8.84	8.60	8.89	8.67
Gumbel	14.8	59.7	5.78	5.62	5.94	5.84

As in the previous section consider the weighted least squares where the choice of weights follows Subsection 3. For this missing structure the assumption of $E(\delta X) = E(\delta)E(X)$ does not hold. The asymptotic variances for partial and full imputation when the errors are standard normally distributed are

$$\begin{aligned} AV_{PI\_WLSbad} &= 997.5 \\ AV_{FI\_WLSbad} &= 998.0 \end{aligned}$$

Again this shows an example where full imputation has more asymptotic variance than partial imputation due to the poor estimation of the parameter ϑ.

C. Exponential missing structure

Now consider an exponential missing structure where

$$\pi(X) = E(\delta|X) = \left(1 + \exp\left\{\frac{X - \mu_X}{\sigma_X}\right\}\right)^{-1}.$$

The plot for the exponential missing structure is given in Figure 5.

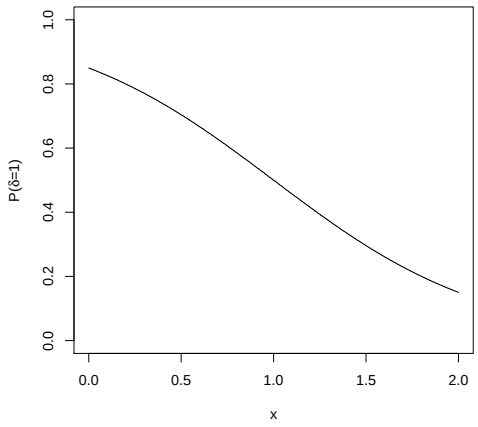

Fig. 5. Exponential missing structure.

As before $E(X) = 1$, $E(X^2) = 4/3$, and $\sigma_X = 1/3$. The other needed pieces are solved as

$$\begin{aligned}
E(\delta) &= \int_0^2 \frac{1}{2}\frac{1}{1+e^{(X-1)\sqrt{3}}}dX = 1/2 \\
E(\delta X) &= \int_0^2 X\frac{1}{2}\frac{1}{1+e^{(X-1)\sqrt{3}}}dX = 0.37355217 \\
E(\delta X^2) &= \int_0^2 X^2\frac{1}{2}\frac{1}{1+e^{(X-1)\sqrt{3}}}dX = 0.4137710 \\
E\left\{\frac{1}{E(\delta|X)}\right\} &= \int_0^2 \frac{1}{2}\left\{1+e^{(X-1)\sqrt{3}}\right\}dX = 2.58059 \\
E\left\{\frac{X^2}{E(\delta|X)}\right\} &= \int_0^2 X^2\frac{1}{2}\left\{1+e^{(X-1)\sqrt{3}}\right\}dX = 5.1455
\end{aligned}$$

Here $E(X) = 1$, $E(X^2) = 4/3$, $\sigma_X = 1/3$, $E(\delta) = 1/2$, $E(\delta X) = 0.37355$, $E(\delta X^2) = 0.41377$, $E\{1/E(\delta|X)\} = 2.58059$, and $E\{X^2/E(\delta|X)\} = 5.1455$. The asymptotic variances for this model are given in Table IV. For this scenario the Listwise Deletion method is not of interest because it is biased. Asymptotically the Listwise Deletion estimator will not converge to the true $E(Y)$. The Propensity Score

method has much higher variance than the imputation methods. As in the previous two sections when the error structure is Uniform the OLS method outperforms the other methods. When the error structure is Normal the OLS and efficient estimates are the same. The best method in every case except the Uniform is full imputation with an efficient estimate of ϑ.

Table IV. The asymptotic variances where the missing structure is exponential.

	LD	PS	PI_EFF	FI_EFF	PI_OLS	FI_OLS
U(-1,1)	–	38.2	4.194	4.194	3.86	3.806
Normal	–	39.9	5.58	5.417	5.58	5.417
DExp	–	42.51	7.138	6.647	8.159	7.834
t_3	–	45.11	9.206	8.47	10.739	10.25
Gamma(2,1)-2	–	42.51	7.138	6.647	8.159	7.834
Logistic	–	45.80	11.2	10.71	11.49	10.95
Gumbel	–	41.55	6.65	6.29	7.24	6.98

As in the previous two sections consider the weighted least squares where the choice of weights follows Subsection 3. For this missing structure the assumption of $E(\delta X) = E(\delta)E(X)$ does not hold. The asymptotic variances for partial and full imputation when the errors are standard normally distributed are

$$\begin{aligned} AV_{PI\_WLSbad} &= 997.7 \\ AV_{FI\_WLSbad} &= 998.0 \end{aligned}$$

Again this shows an example where full imputation has more asymptotic variance than partial imputation due to the poor estimation of the parameter ϑ.

D. Simulation results with finite sample sizes

This section uses simulations to compare Full Imputation, Partial Imputation, and the Propensity Score method. The simulated Mean Square Error of the estimators for $E(Y)$ is shown for various estimates of the parameter ϑ. OLS denotes the Ordinary Least Squares estimate, OSI denotes to the One Step Improvement estimator, and MELE denotes the Maximum Empirical Likelihood Estimator with one, two or three constraints which are called MELE1, MELE2, and MELE3, respectively. The results are from 20,000 simulations.

Figure 6 shows results with no missing data when the errors are normally distributed. Because there is no missing data the Partial Imputation method is simply the empirical estimator $\bar{Y}$ and therefore does not depend on ϑ. Since the errors are normal Full imputation is asymptotically efficient for all five estimates of ϑ. For small sample sizes the OSI method has a larger MSE, partly due to the difficulty of estimating the score function. The MSE for the Propensity Score method is given below each plot, which is again the MSE of $\bar{Y}$.

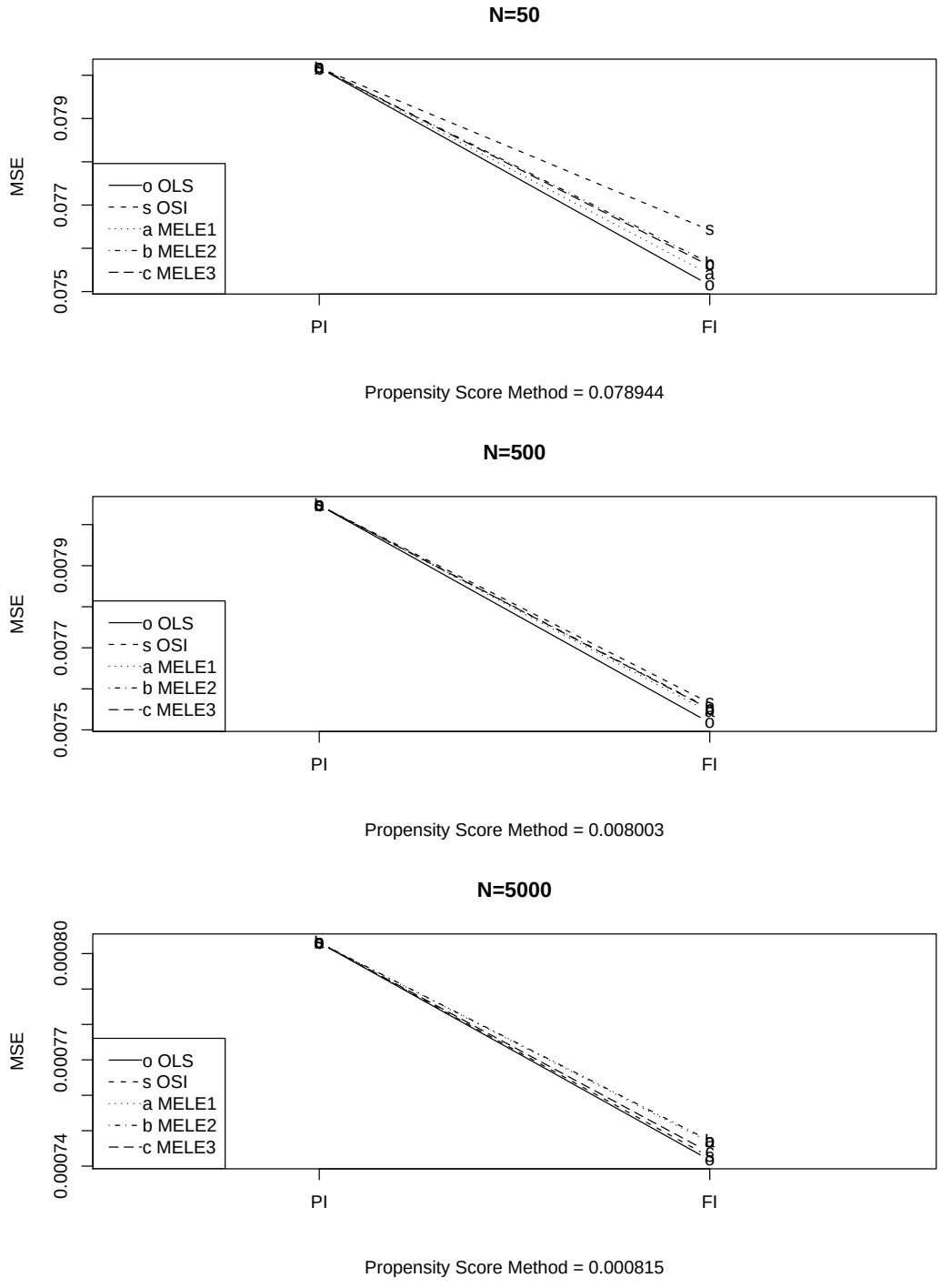

Fig. 6. MSE for estimating $E[Y]$ under normal errors with no missing data

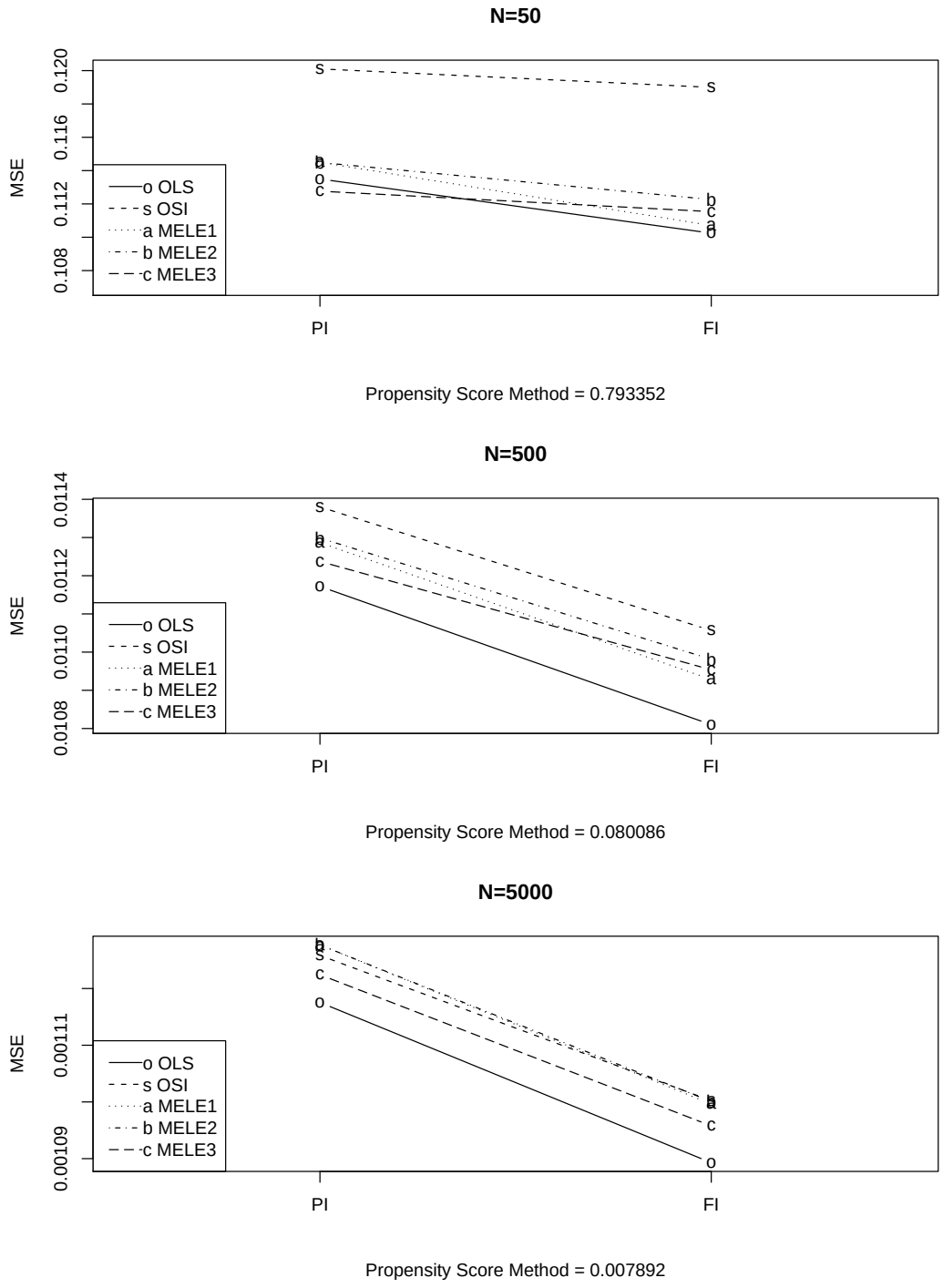

Fig. 7. MSE for estimating $E[Y]$ under normal errors with an exponential missing structure

Figure 7 shows results from an exponential missing structure. With missing data the variance of the partial imputation method depends on the estimate of ϑ. The errors are again generated from the normal distribution so Full Imputation is efficient with any estimate of ϑ. With small sample sizes the OLS estimate has the smallest MSE and OSI has a larger MSE. The MELE3 method performs better than MELE1 or MELE2 for the larger sample size, which is expected. In every case Full Imputation is better than Partial Imputation.

Figure 8 shows results with gamma errors and an exponential missing stucture. In this case for small sample sizes there is a dramatic difference in the MSE for estimating $E(Y)$ depending on the estimate of ϑ. The OSI estimate cannot estimate the score function well for small sample sizes, but for a large sample size it performs very well. The OLS estimate has more variance, and that difference grows as the sample size increases.

For graphs of other error distributions across the different structures of the missing data, as well as tables showing the MSE values see Appendix A.

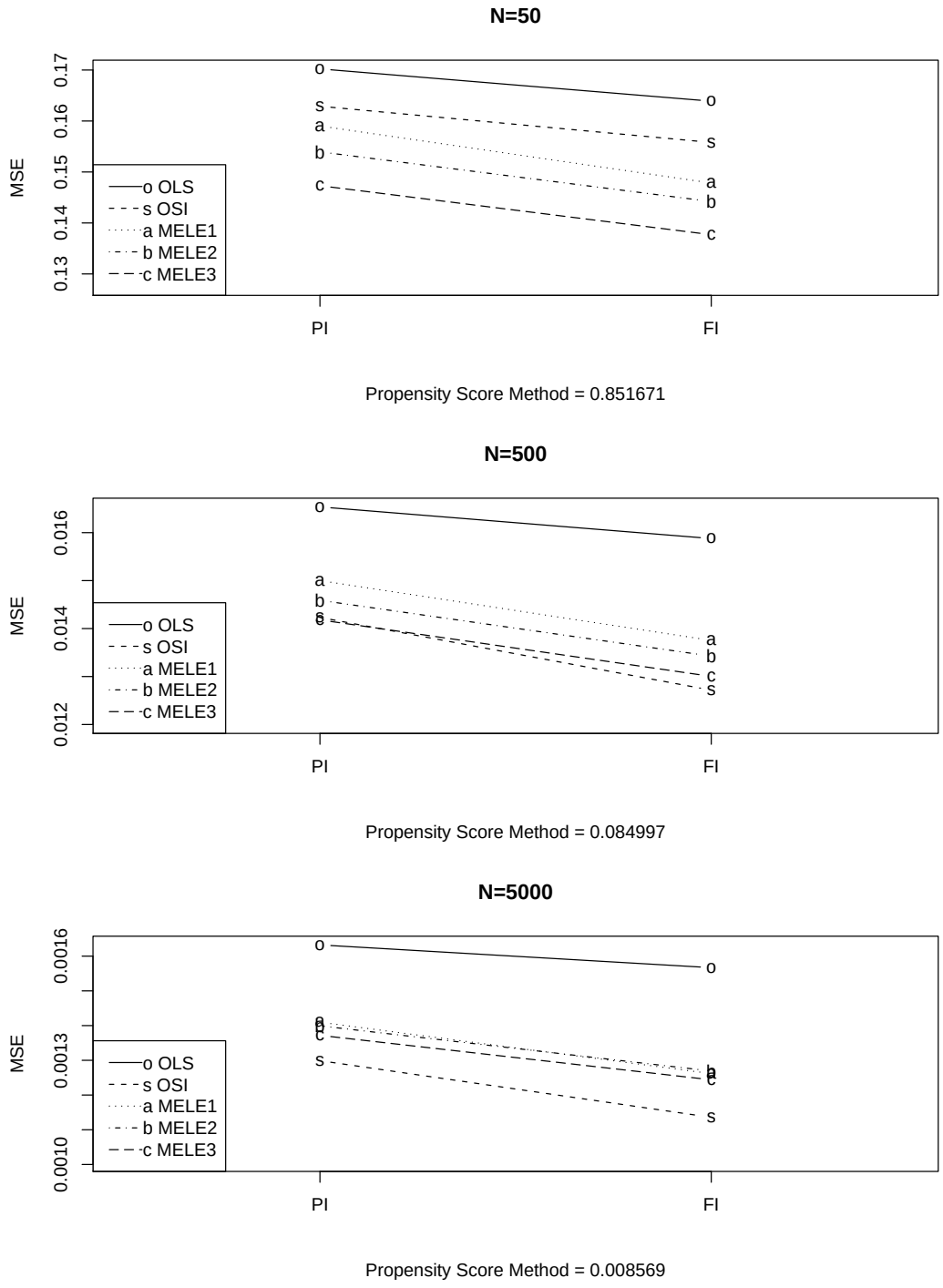

Fig. 8. MSE for estimating $E[Y]$ under gamma errors with an exponential missing structure

CHAPTER VI

SUMMARY

When the unknown error distribution is in fact normal the Ordinary Least Squares estimator is efficient for estimating the parameter in linear regression. If the errors are not normal, then a One Step Improvement estimate or a Maximum Empirical Likelihood estimate can be used to estimate the parameter efficiently.

When estimating the mean response Listwise Deletion can be biased depending on the missing structure. The Propensity Score method is an improvement as it is unbiased, but neither Listwise Deletion nor the Propensity Score method are efficient. Estimating the mean response efficiently requires imputation. For both Partial Imputation and Full Imputation the performance depends on the estimate of the regression parameter, but in general Full Imputation is better than Partial Imputation. Only when the parameter is estimated very poorly will Partial Imputation have less variance than full imputation. The efficient estimate for the mean response is Full Imputation with an efficient estimate of the parameter.

REFERENCES

Andrews, D. W. K. and D. Pollard (1994). An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review 62*, 119–132. 49

Bell, B. A., J. D. Kromrey, and J. M. Ferron (2009). Missing data and complex samples: The impact of listwise deletion vs. subpopulation analysis on statistical bias and hypothesis test results when data are mcar and mar. *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*. 26

Dong, W. and X. C. Song (2009). Empirical likelihood for estimating equation with missing values. *Annals of Statistics 37*, 490–517. 25

Elliot, M. (2008). Healthy for life: Accounting for transcription errors using multiple imputation. *CHANCE 21*, 14–23. 1

Forrester, J., W. Hooper, H. Peng, and A. Schick (2003). On the construction of efficient estimators in semiparametric models. *Statistical Decisions 21*, 109–138. 2, 16

Koul, H. L., U. U. Müller, and A. Schick (2012). Complete case analysis revisited. Draft. For more information contact uschi@stat.tamu.edu. 17

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2^{nd} ed.). Hoboken, NJ: Wiley-Interscience. 1

Müller, U. U. (2007). Weighted least squares estimators in possibly misspecified nonlinear regression. *Metrika 66*, 39–59. 47, 48

Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Stat. 37*, 2245–2277. 1, 5, 40, 44, 48

Müller, U. U., A. Schick, and W. Wefelmeyer (2004). Estimating linear functionals of the error distribution in nonparametric regression. *Journal of Statistical Planning and Inference 119*, 75–93. 49

Müller, U. U., A. Schick, and W. Wefelmeyer (2006). *Probability Statistics and Modeling in Public Health - Symposium in Honor of Marvin Zelen*, pp. 350–363. New York, NY: Springer. 1, 4

Müller, U. U., A. Schick, and W. Wefelmeyer (2007). Estimating the error distribution function in semiparametric regression. *Statistics & Decisions 25*, 1–18. 158

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrica 75*, 237–249. 17

Owen, A. B. (2001). *Empirical Likelihood*. Boca Raton, FL: Chapman & Hall/CRC. 18, 41

Peng, H. and A. Schick (2012). An empirical likelihood approach to goodness of fit testing. Draft. For more information contact anton@math.binghamton.edu. 2, 17

Rosenbaum, R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55. 25

Schick, A. (1993). On efficient estimation in regression models. *Ann. Stat. 21*, 1486–1521. 1

Schick, A. (1996a). Root-n consistent and efficient estimation in semiparametric additive regression models. *Statistics and Probability Letters 30*, 45–51. 28, 160

Schick, A. (1996b). Root-n consistent estimation in partly linear regression models. *Statistics and Probability Letters 28*, 353–358. 28

Schick, A. (1996c). Weighted least squares estimates in partly linear regression models. *Statistics and Probability Letters 27*, 281–287. 50

Wang, Q., O. Linton, and W. Härdle (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association 99*, 334–345. 158

Zhi, D. (2012). Univariate kernel density estimation. Webpage describing method: http://www.stat.duke.edu/∼zo2/shared/research/readings/kernelsmoothing.pdf. 17

APPENDIX

A. Additional results and tables of MSE

1. Estimation of ϑ

Using the model as defined on page 18 where $\vartheta = 3$, $E[\epsilon] = 0$, $X \sim \text{Uniform}(0, 2)$, with ϵ and X independent. Then $\hat{\vartheta}$ was calculated using each of the following methods:

1. OLS: Ordinary Least Squares. Estimator will be efficient when the unknown error is in fact normally distributed.

2. OSI: One Step Improvement. Estimator is asymptotically efficient.

3. MELE1: Maximum Empirical Likelihood Estimator with one constraint. For small sample sizes one constraint could be sufficient to acheive efficiency.

4. MELE2: Maximum Empirical Likelihood Estimator with two constraints on the basis. The extra constraint handles larger sample sizes.

5. MELE3: Maximum Empirical Likelihood Estimator with three constraints on the basis. The larger the sample size the more constraints needed to achieve efficiency.

The Mean Square Error (MSE) was calculated for each simulation using various methods of estimating ϑ. Figure 9 shows the MSE for estimating ϑ when the error are normally distributed. In such a case all the estimators are efficient, and they appear to be very similar. The OSI method needs to estimate the score function, and the Maximum Emperical Likelihood methods use a grid search to find the estimate for ϑ, so the smallest MSE comes from the OLS method. When the errors follow the standard logistic distribution as in Figure 10, more separation can be seen between the methods. OSI happens to do poorly for small sample sizes due to the estimation of the score function, while MELE with one constraint does well.

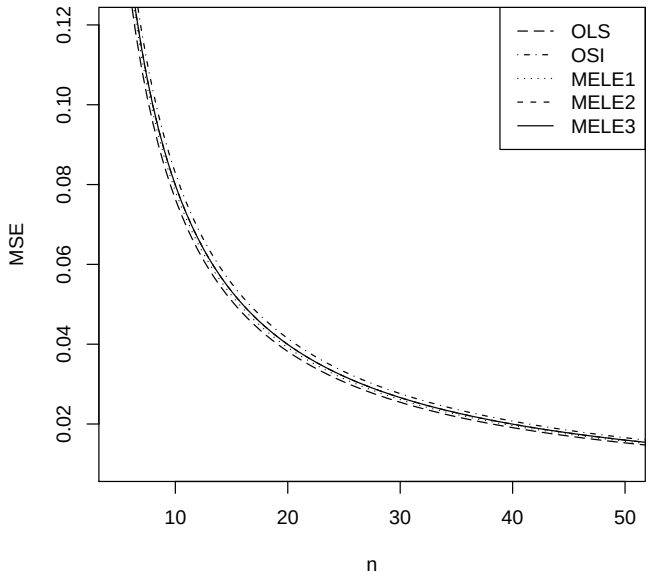

Fig. 9. MSE for various methods of estimating ϑ under normal errors.

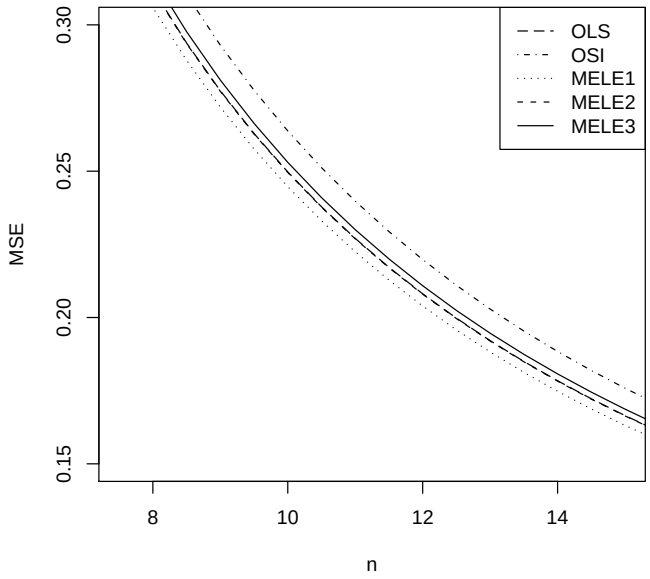

Fig. 10. MSE for various methods of estimating ϑ under logistic errors.

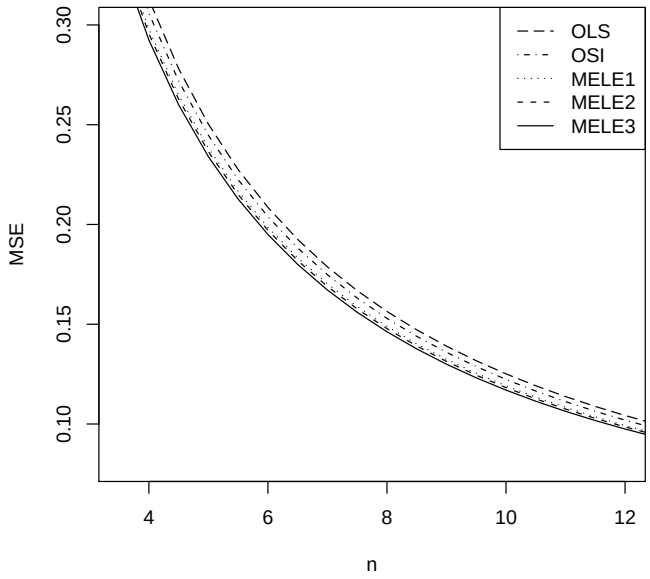

Fig. 11. MSE for various methods of estimating ϑ under Gumbel errors.

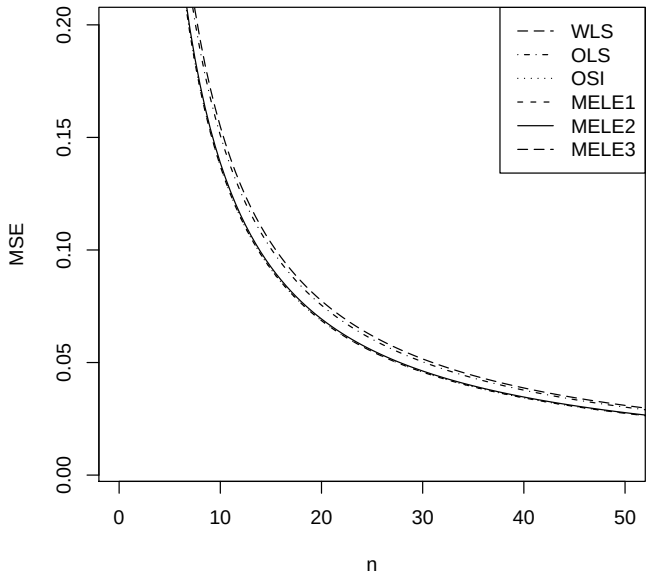

Fig. 12. MSE for various methods of estimating ϑ under double exponential errors.

The results from the Gumbel distribution are in Figure 11. The OLS is clearly the least desirable method, while the MELE with three constraints performs the best. For the double exponential model shown in Figure 12 the OLS and OSI methods struggle while the three MELE methods are very close. The results of all these models show that the OLS works well with distributions that are similar to the normal, but departures from normality make the MELE methods preferable. For small sample sizes the OSI method often has more variability than the MELE methods.

Table V shows the values of the MSE averaged over the 20,000 iterations. The simulation used three sample sizes, $N = 50$, $N = 500$, and $N = 5000$. Then the observed values of MSE were fit to a $1/N$ model for the plots shown in the paper.

2. Estimation of $E(Y)$

Using the model as defined in Chapter V where the goal is to estimate $E(Y)$ under the simple linear model where $Y = \vartheta X + \epsilon$. Let $X \sim \ Uniform(0, 2)$, and $\vartheta = 3$. The following error distributions of ϵ are considered:

1. $\epsilon \sim \ Normal(0, 1)$ where the OLS method is efficient;
2. $\epsilon \sim \ t$ with 3 degrees of freedom;
3. $\epsilon \sim \ Gamma$ with a variance of 2 shifted to have a mean of 0.
4. $\epsilon \sim \ Logistic$ which under the standard distribution has heavy tails compared to the Normal distribution.
5. $\epsilon \sim \ Gumbel$ which under the standard distribution is similar to the Normal but is skewed.

For each scenario the asymptotic variances will be compared using the following methods:

Table V. Simulation results for the estimation of ϑ from 20,000 iterations.

		Normal errors	t_2 errors	Logistic errors	Gumbel errors	Gamma errors	DExp errors
N=50	OLS	0.01527	0.1989	0.04995	0.02503	0.03085	0.03097
	OSI	0.01658	0.07616	0.05278	0.02449	0.02661	0.03022
	MELE1	0.01567	0.05795	0.04895	0.02384	0.02772	0.02747
	MELE2	0.01598	0.05785	0.04992	0.02366	0.02662	0.02746
	MELE3	0.01598	0.06247	0.05062	0.02341	0.02491	0.02776
N=500	OLS	0.001499	0.02289	0.004972	0.002494	0.002986	0.002991
	OSI	0.001542	0.007142	0.005023	0.002245	0.002068	0.002714
	MELE1	0.001521	0.006946	0.004846	0.002329	0.002459	0.002557
	MELE2	0.001527	0.006969	0.004859	0.002278	0.002383	0.002562
	MELE3	0.001523	0.006941	0.004898	0.002258	0.002254	0.002535
N=5000	OLS	0.000147	0.002705	0.0004896	0.0002462	0.0003038	0.0002969
	OSI	0.0001487	0.0008128	0.000484	0.0002184	0.0001849	0.0002535
	MELE1	0.0001504	0.000724	0.0004759	0.0002281	0.0002284	0.0002523
	MELE2	0.0001505	0.0007245	0.0004769	0.0002228	0.0002314	0.0002523
	MELE3	0.0001486	0.0007121	0.0004774	0.0002203	0.0002268	0.0002463

1. LD - Listwise Deletion (Equation 3.5);
2. PS - Propensity Score using the true $\pi(X)$ (Equation 3.7);
3. PI_EFF - Partial Imputation with an efficient estimate (Equation 4.3);
4. FI_EFF - Full imputation with an efficient estimate (Equation 4.4);
5. PI_OLS - Partial Imputation using Ordinary Least Squares (Equation 4.15);
6. FI_OLS - Full Imputation using Ordinary Least Squares (Equation 4.16).

There are three different types of missing structures that are considered:

1. None - When there is no missing data;
2. Gaussian - a normal probability of the response being missing as defined in Section B;
3. Exponential - where the probability of the response being missing is higher on one end as defined in Section C.

a. No missing data

The graph for the normally distributed data is shown on page 69. Figure 13 shows the MSE for the estimates of $E(Y)$ when the errors have a t distribution. Figure 14 shows the same for when the errors have a gamma distribution. Figure 15 is for the logistic distribution, and Figure 16 is for the Gumbel distribution.

Table VI shows the MSE values from 20,000 simulations where there is no missing data and the errors are normally distributed. Table VII shows the same thing for errors that have a t distribution, while Table VIII, Table IX, and Table X have errors that are logistic, Gumbel, and gamma respectively.

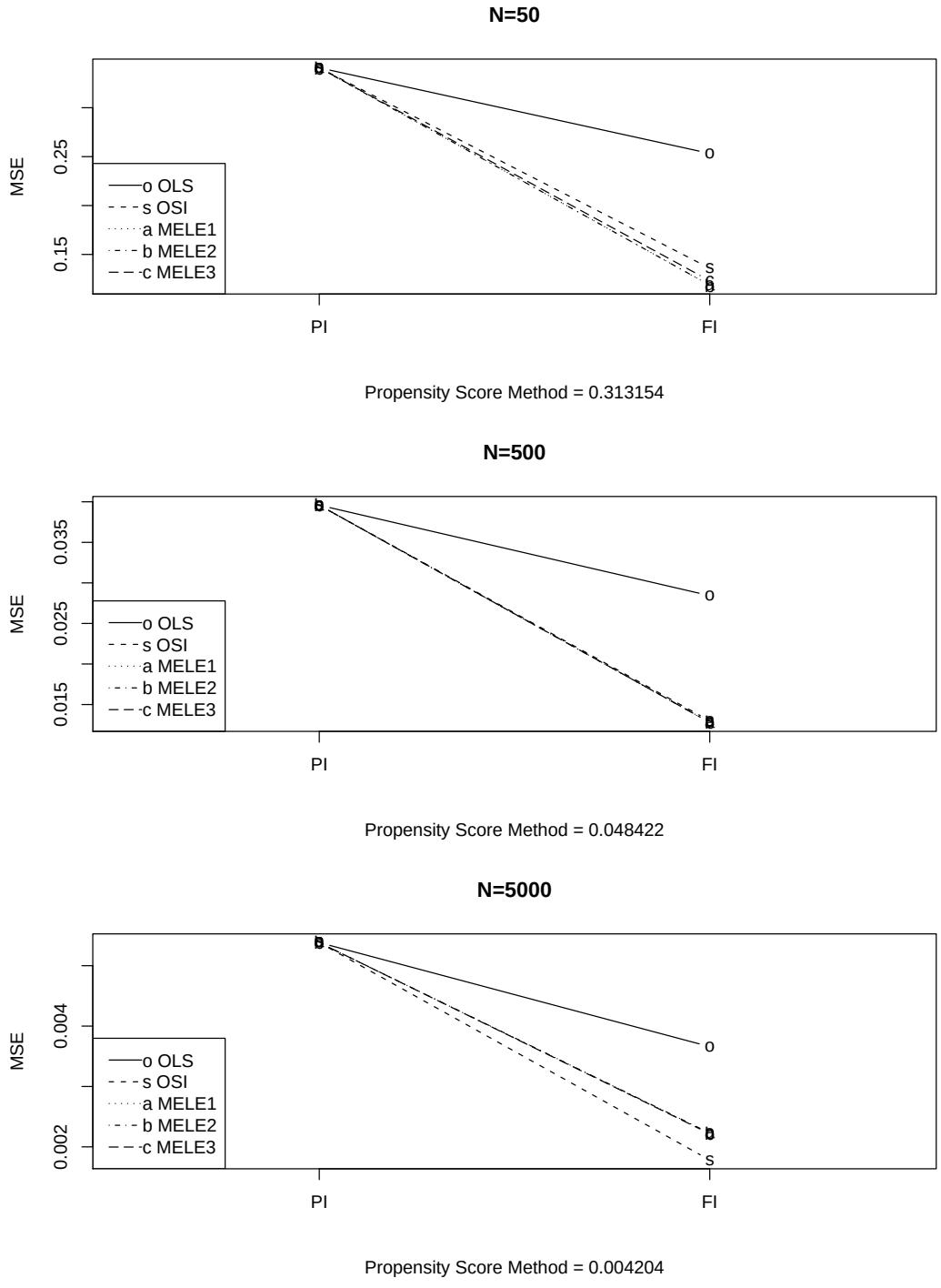

Fig. 13. MSE for estimating $E[Y]$ where the errors have the t distribution and no missing data

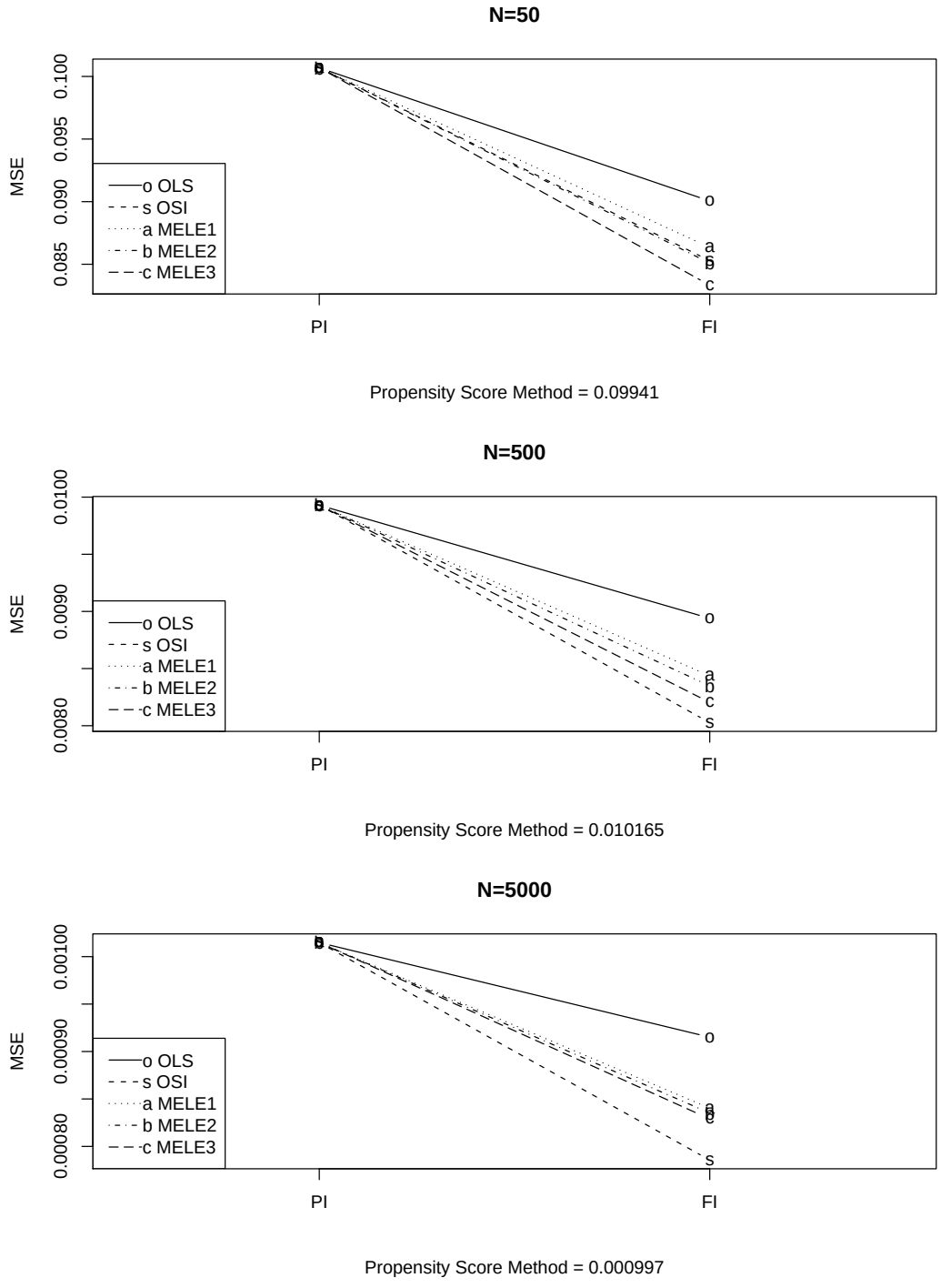

Fig. 14. MSE for estimating $E[Y]$ where the errors have the gamma distribution and no missing data

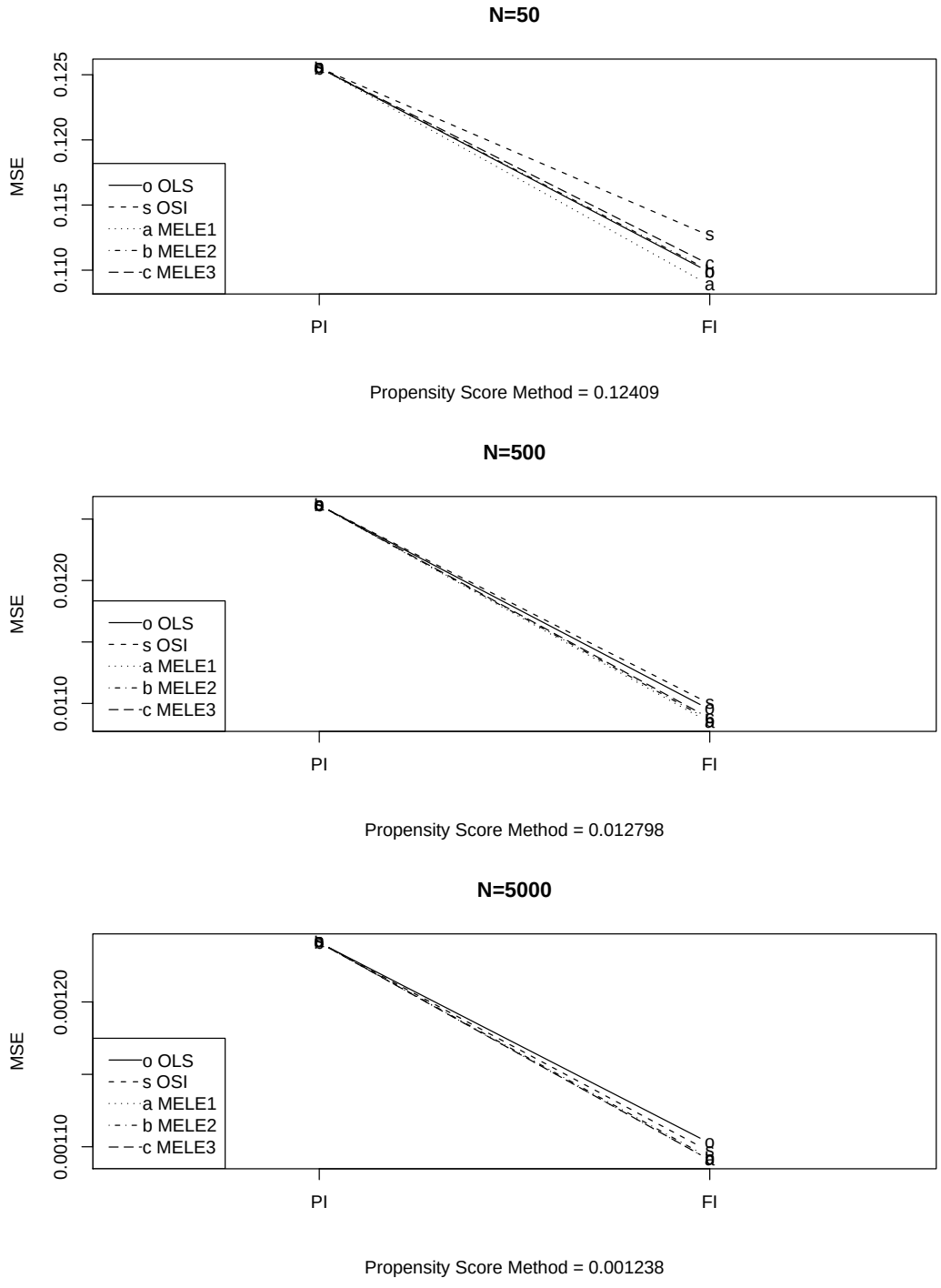

Fig. 15. MSE for estimating $E[Y]$ where the errors have the logistic distribution and no missing data

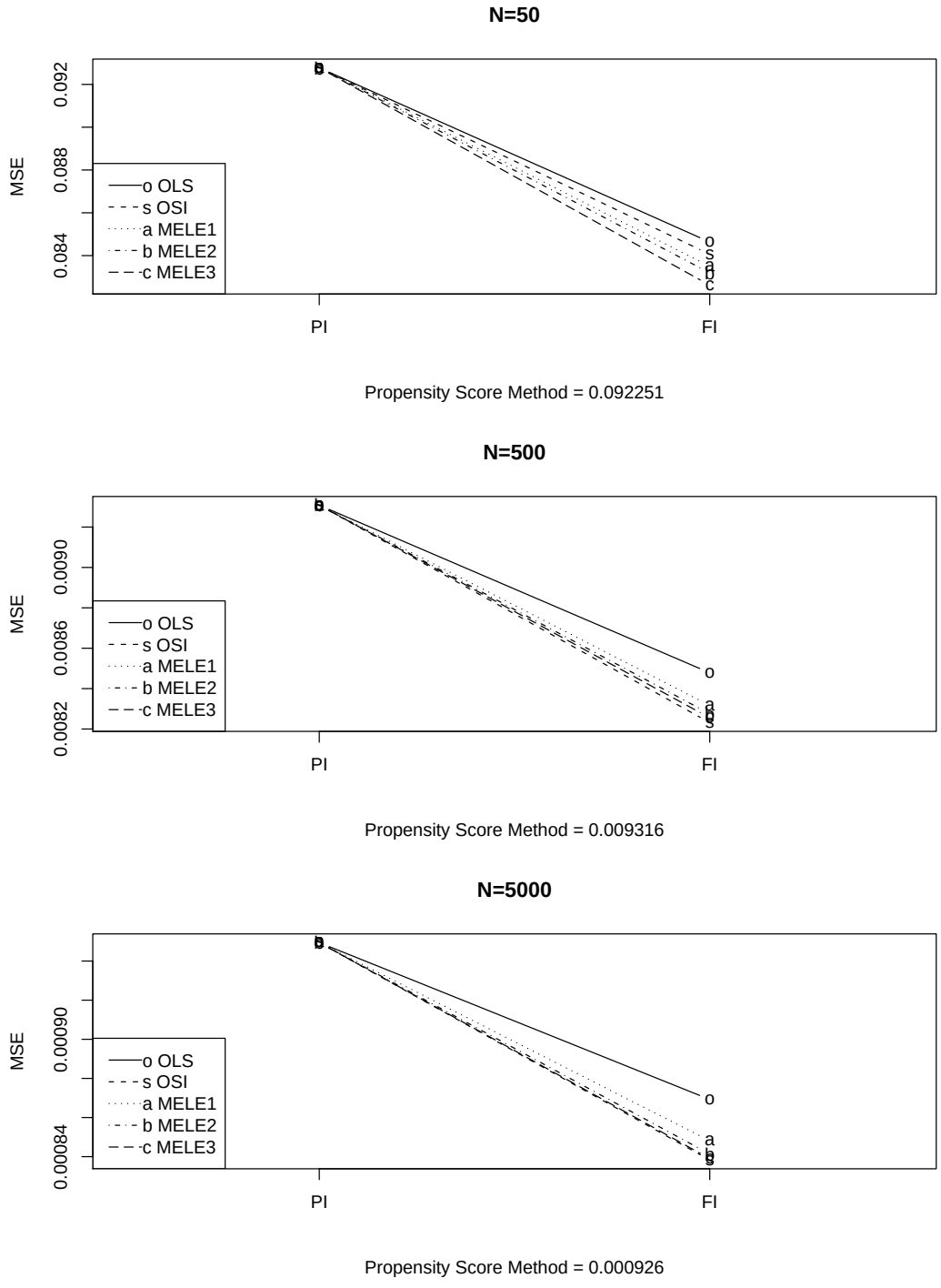

Fig. 16. MSE for estimating $E[Y]$ where the errors have the Gumbel distribution and no missing data

Table VI. Simulation results showing the MSE for the estimation of $E[Y]$where there is no missing data and the errors have a normal distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.08017	0.08017	0.07514
	OSI	0.08017	0.08017	0.07642
	MELE1	0.08017	0.08017	0.0754
	MELE2	0.08017	0.08017	0.07566
	MELE3	0.08017	0.08017	0.0756
N=500	OLS	0.008047	0.008047	0.007517
	OSI	0.008047	0.008047	0.007565
	MELE1	0.008047	0.008047	0.007543
	MELE2	0.008047	0.008047	0.007552
	MELE3	0.008047	0.008047	0.00755
N=5000	OLS	0.0008031	0.0008031	0.0007417
	OSI	0.0008031	0.0008031	0.0007426
	MELE1	0.0008031	0.0008031	0.0007467
	MELE2	0.0008031	0.0008031	0.0007472
	MELE3	0.0008031	0.0008031	0.0007438

Table VII. Simulation results showing the MSE for the estimation of $E[Y]$where there is no missing data and the errors have a t_2 distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.3408	0.3408	0.2533
	OSI	0.3408	0.3408	0.1359
	MELE1	0.3408	0.3408	0.1185
	MELE2	0.3408	0.3408	0.1185
	MELE3	0.3408	0.3408	0.123
N=500	OLS	0.03958	0.03958	0.02842
	OSI	0.03958	0.03958	0.01299
	MELE1	0.03958	0.03958	0.01276
	MELE2	0.03958	0.03958	0.01279
	MELE3	0.03958	0.03958	0.01279
N=5000	OLS	0.005383	0.005383	0.003659
	OSI	0.005383	0.005383	0.001781
	MELE1	0.005383	0.005383	0.002227
	MELE2	0.005383	0.005383	0.002228
	MELE3	0.005383	0.005383	0.002217

Table VIII. Simulation results showing the MSE for the estimation of $E[Y]$where there is no missing data and the errors have a logistic distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.1255	0.1255	0.1098
	OSI	0.1255	0.1255	0.1126
	MELE1	0.1255	0.1255	0.1088
	MELE2	0.1255	0.1255	0.11
	MELE3	0.1255	0.1255	0.1104
N=500	OLS	0.01261	0.01261	0.01095
	OSI	0.01261	0.01261	0.011
	MELE1	0.01261	0.01261	0.01084
	MELE2	0.01261	0.01261	0.01086
	MELE3	0.01261	0.01261	0.01088
N=5000	OLS	0.001241	0.001241	0.001103
	OSI	0.001241	0.001241	0.001097
	MELE1	0.001241	0.001241	0.001091
	MELE2	0.001241	0.001241	0.001092
	MELE3	0.001241	0.001241	0.001091

Table IX. Simulation results showing the MSE for the estimation of $E[Y]$where there is no missing data and the errors have a Gumbell distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.09278	0.09278	0.08463
	OSI	0.09278	0.09278	0.08406
	MELE1	0.09278	0.09278	0.08349
	MELE2	0.09278	0.09278	0.08319
	MELE3	0.09278	0.09278	0.08261
N=500	OLS	0.009308	0.009308	0.008478
	OSI	0.009308	0.009308	0.008231
	MELE1	0.009308	0.009308	0.008317
	MELE2	0.009308	0.009308	0.008275
	MELE3	0.009308	0.009308	0.008256
N=5000	OLS	0.0009494	0.0009494	0.0008692
	OSI	0.0009494	0.0009494	0.0008382
	MELE1	0.0009494	0.0009494	0.0008481
	MELE2	0.0009494	0.0009494	0.0008413
	MELE3	0.0009494	0.0009494	0.000839

Table X. Simulation results showing the MSE for the estimation of $E[Y]$where there is no missing data and the errors have a gamma distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.1007	0.1007	0.09005
	OSI	0.1007	0.1007	0.08528
	MELE1	0.1007	0.1007	0.08634
	MELE2	0.1007	0.1007	0.08509
	MELE3	0.1007	0.1007	0.08332
N=500	OLS	0.009929	0.009929	0.008939
	OSI	0.009929	0.009929	0.008027
	MELE1	0.009929	0.009929	0.008436
	MELE2	0.009929	0.009929	0.008348
	MELE3	0.009929	0.009929	0.008204
N=5000	OLS	0.001015	0.001015	0.0009151
	OSI	0.001015	0.001015	0.0007856
	MELE1	0.001015	0.001015	0.0008403
	MELE2	0.001015	0.001015	0.0008351
	MELE3	0.001015	0.001015	0.0008289

b. Gaussian missing structure

Figure 17 shows the MSE for the estimates of $E(Y)$ when the errors have a normal distribution. Figure 18 shows the same for when the errors have a t distribution. Figure 19 is for the gamma distribution, Figure 20 is for the logistic distribution, and Figure 21 is for the Gumbel distribution.

Table XI shows the MSE values from 20,000 simulations where the missing structure is Gaussian and the errors are normally distributed. Table XII shows the same thing for errors that have a t distribution, while Table XIII, Table XIV, and Table XV have errors that are logistic, Gumbel, and gamma respectively.

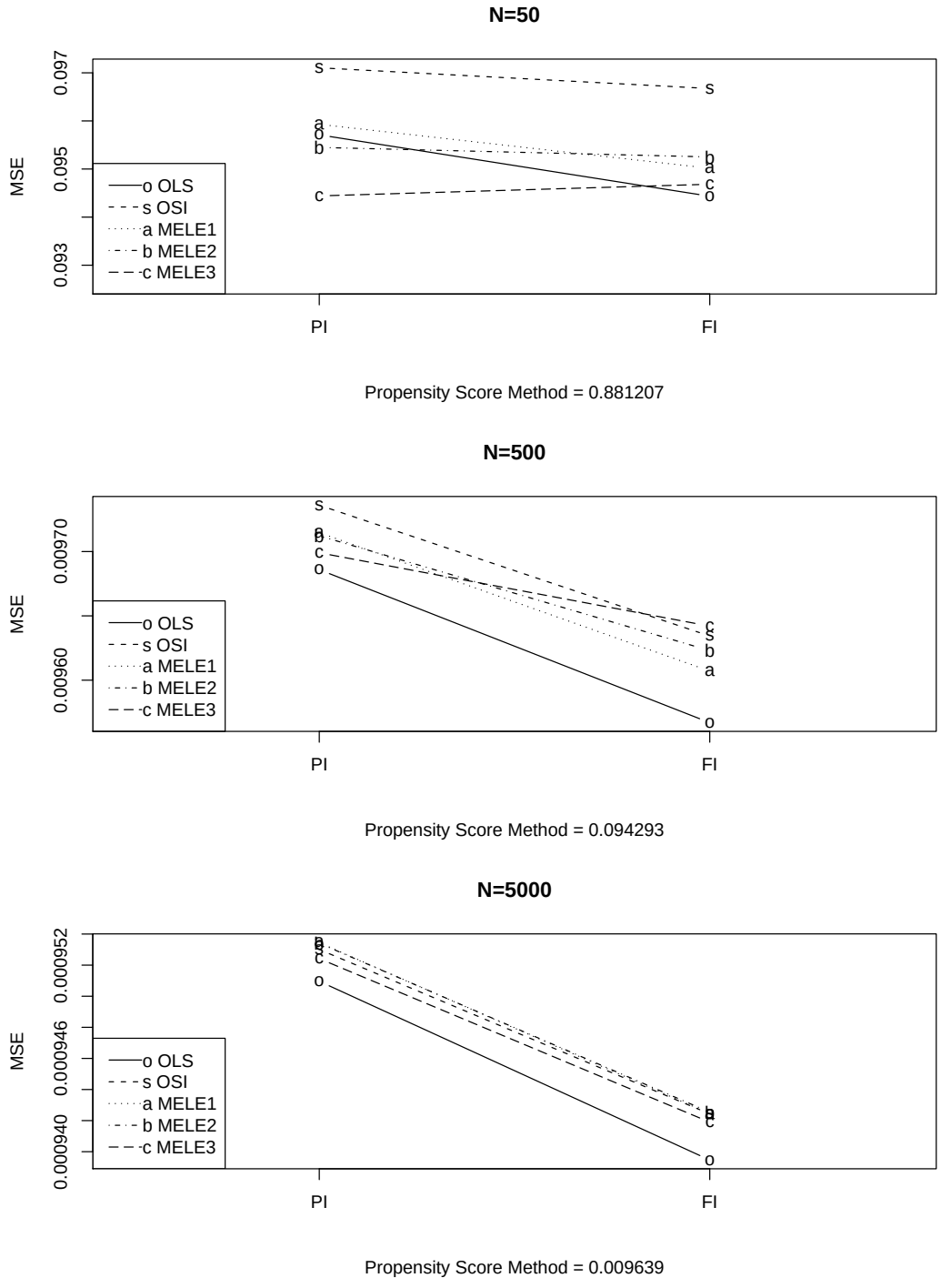

Fig. 17. MSE for estimating $E[Y]$ where the errors have the normal distribution and a Gaussian missingness structure

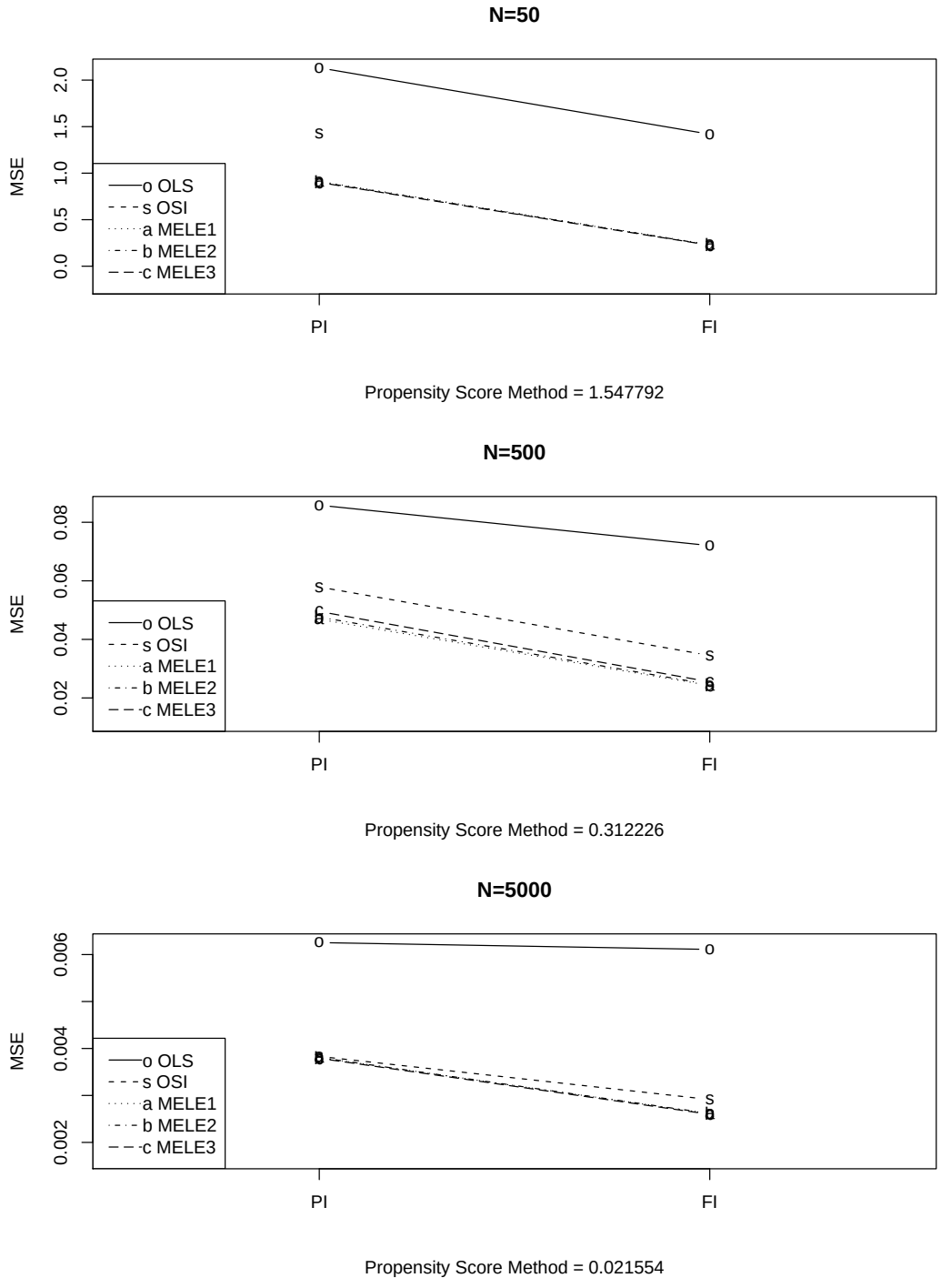

Fig. 18. MSE for estimating $E[Y]$ where the errors have the t distribution and a Gaussian missingness structure

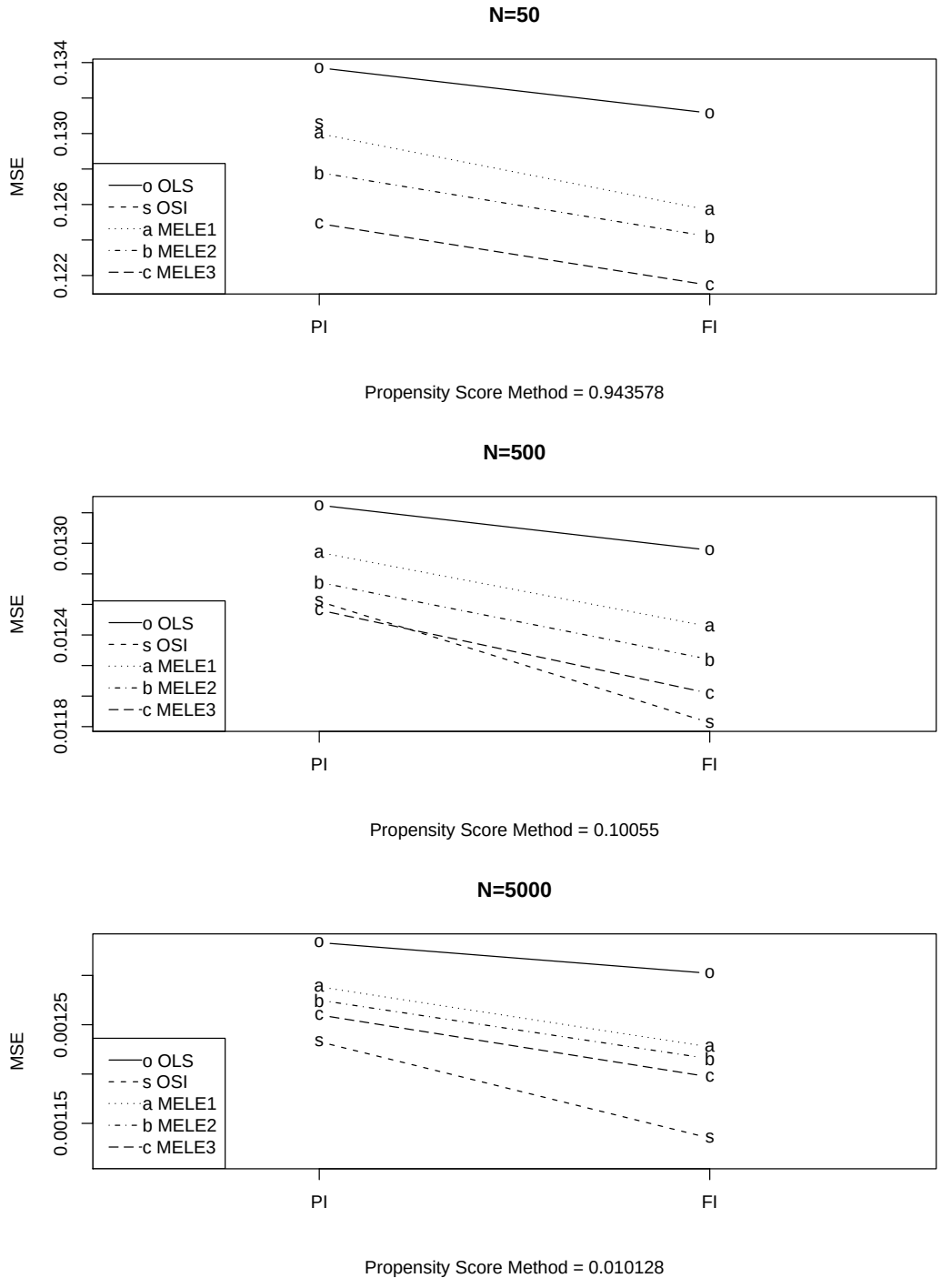

Fig. 19. MSE for estimating $E[Y]$ where the errors have the gamma distribution and a Gaussian missingness structure

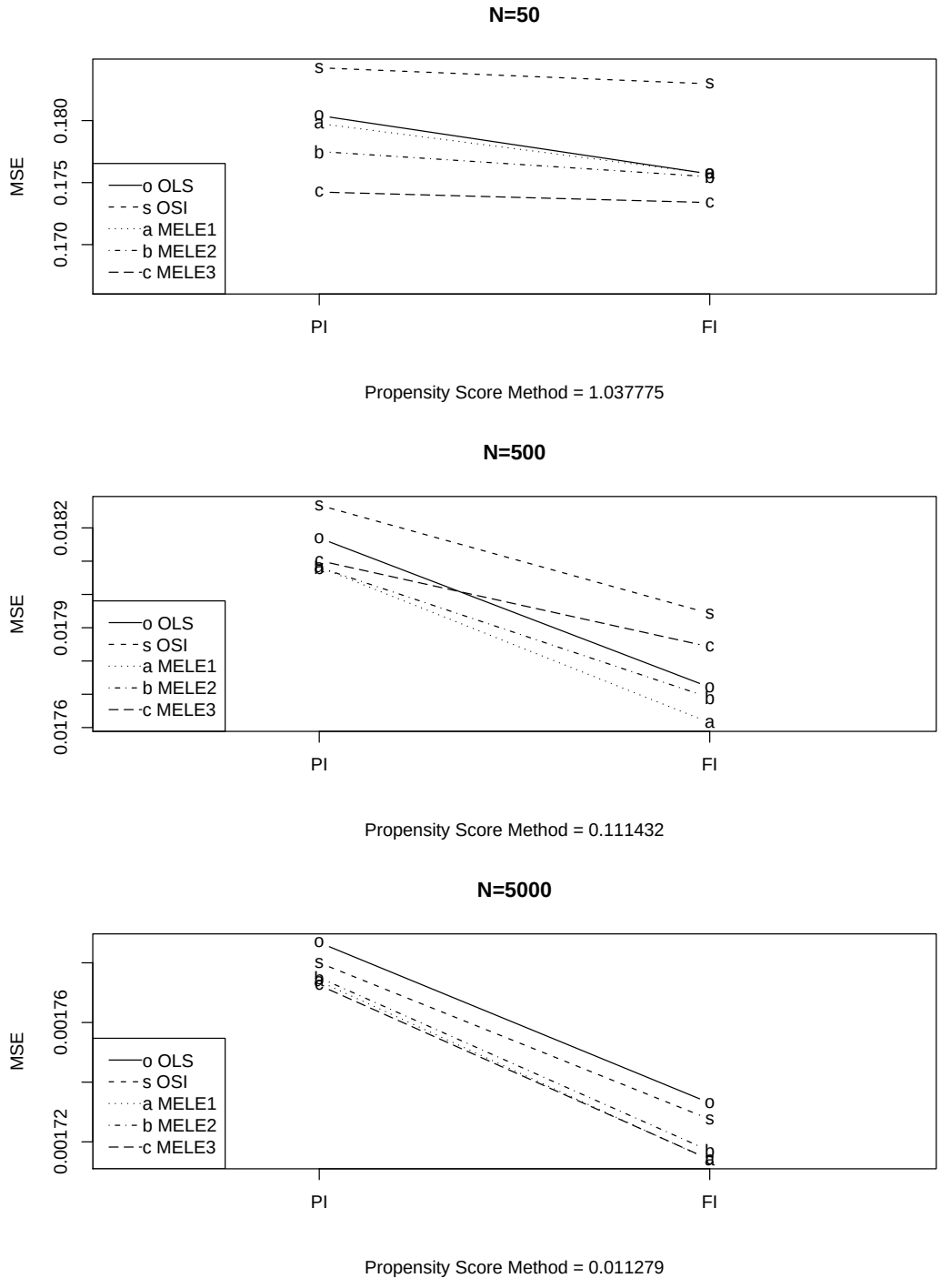

Fig. 20. MSE for estimating $E[Y]$ where the errors have the logistic distribution and a Gaussian missingness structure

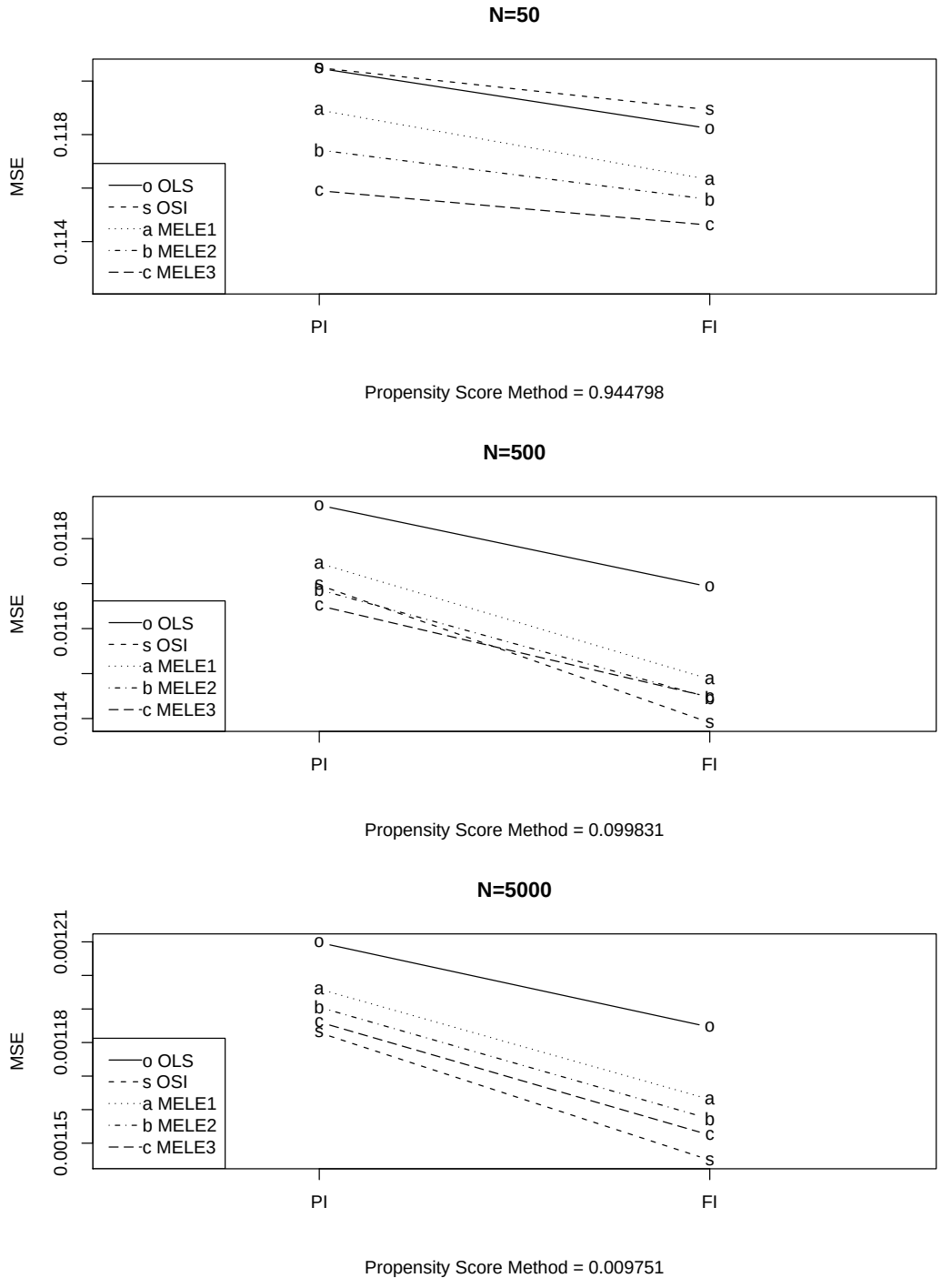

Fig. 21. MSE for estimating $E[Y]$ where the errors have the Gumbel distribution and a Gaussian missingness structure

Table XI. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is Gaussian and the errors have a normal distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.1777	0.09571	0.09443
	OSI	0.1777	0.0971	0.09667
	MELE1	0.1777	0.09593	0.09502
	MELE2	0.1777	0.09545	0.09525
	MELE3	0.1777	0.09444	0.09468
N=500	OLS	0.01806	0.009686	0.009567
	OSI	0.01806	0.009736	0.009634
	MELE1	0.01806	0.009715	0.009607
	MELE2	0.01806	0.009712	0.009623
	MELE3	0.01806	0.009699	0.009642
N=5000	OLS	0.001776	0.000951	0.0009395
	OSI	0.001776	0.000953	0.0009424
	MELE1	0.001776	0.0009535	0.0009424
	MELE2	0.001776	0.0009534	0.0009425
	MELE3	0.001776	0.0009524	0.0009419

Table XII. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is Gaussian and the errors have a t_2 distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	3.387	2.132	1.418
	OSI	3.387	1.431	NaN
	MELE1	3.387	0.9123	0.2325
	MELE2	3.387	0.9055	0.2284
	MELE3	3.387	0.9005	0.2254
N=500	OLS	0.115	0.08583	0.07203
	OSI	0.115	0.05789	0.03451
	MELE1	0.115	0.04686	0.02426
	MELE2	0.115	0.04762	0.02459
	MELE3	0.115	0.04953	0.02549
N=5000	OLS	0.007591	0.006255	0.006109
	OSI	0.007591	0.003831	0.002914
	MELE1	0.007591	0.003808	0.002613
	MELE2	0.007591	0.003808	0.002616
	MELE3	0.007591	0.003791	0.002598

Table XIII. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is Gaussian and the errors have a logistic distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.2719	0.1804	0.1757
	OSI	0.2719	0.1843	0.183
	MELE1	0.2719	0.1798	0.1757
	MELE2	0.2719	0.1775	0.1755
	MELE3	0.2719	0.1742	0.1734
N=500	OLS	0.02746	0.01817	0.01772
	OSI	0.02746	0.01827	0.01794
	MELE1	0.02746	0.01808	0.01761
	MELE2	0.02746	0.01808	0.01769
	MELE3	0.02746	0.0181	0.01784
N=5000	OLS	0.002713	0.001787	0.001733
	OSI	0.002713	0.00178	0.001728
	MELE1	0.002713	0.001774	0.001714
	MELE2	0.002713	0.001775	0.001717
	MELE3	0.002713	0.001772	0.001714

Table XIV. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is Gaussian and the errors have a Gumbel distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.2059	0.1205	0.1182
	OSI	0.2059	0.1205	0.1189
	MELE1	0.2059	0.1189	0.1163
	MELE2	0.2059	0.1174	0.1156
	MELE3	0.2059	0.1159	0.1146
N=500	OLS	0.02063	0.01187	0.01169
	OSI	0.02063	0.0117	0.01139
	MELE1	0.02063	0.01175	0.01149
	MELE2	0.02063	0.01169	0.01145
	MELE3	0.02063	0.01165	0.01145
N=5000	OLS	0.002079	0.00121	0.001185
	OSI	0.002079	0.001183	0.001145
	MELE1	0.002079	0.001196	0.001163
	MELE2	0.002079	0.001191	0.001157
	MELE3	0.002079	0.001186	0.001152

Table XV. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is Gaussian and the errors have a gamma distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.2193	0.1337	0.1312
	OSI	0.2193	0.1306	0.06195
	MELE1	0.2193	0.13	0.1257
	MELE2	0.2193	0.1278	0.1242
	MELE3	0.2193	0.1249	0.1214
N=500	OLS	0.02198	0.01325	0.01295
	OSI	0.02198	0.01262	0.01183
	MELE1	0.02198	0.01294	0.01245
	MELE2	0.02198	0.01274	0.01224
	MELE3	0.02198	0.01256	0.01202
N=5000	OLS	0.002214	0.001333	0.001302
	OSI	0.002214	0.001233	0.001135
	MELE1	0.002214	0.001288	0.001228
	MELE2	0.002214	0.001275	0.001215
	MELE3	0.002214	0.00126	0.001197

c. Exponential missing structure

The graph for the normally distributed data is shown on page 70, and the graph for the errors with the gamma distribution is on page 72. Figure 13 shows the MSE for the estimates of $E(Y)$ when the errors have a t distribution. Figure 15 is for the logistic distribution, and Figure 16 is for the Gumbel distribution.

Table XVI shows the MSE values from 20,000 simulations where the missing structure is exponential and the errors are normally distributed. Table XVII shows the same thing for errors that have a t distribution, while Table XVIII, Table XIX, and Table XX have errors that are logistic, Gumbel, and gamma respectively.

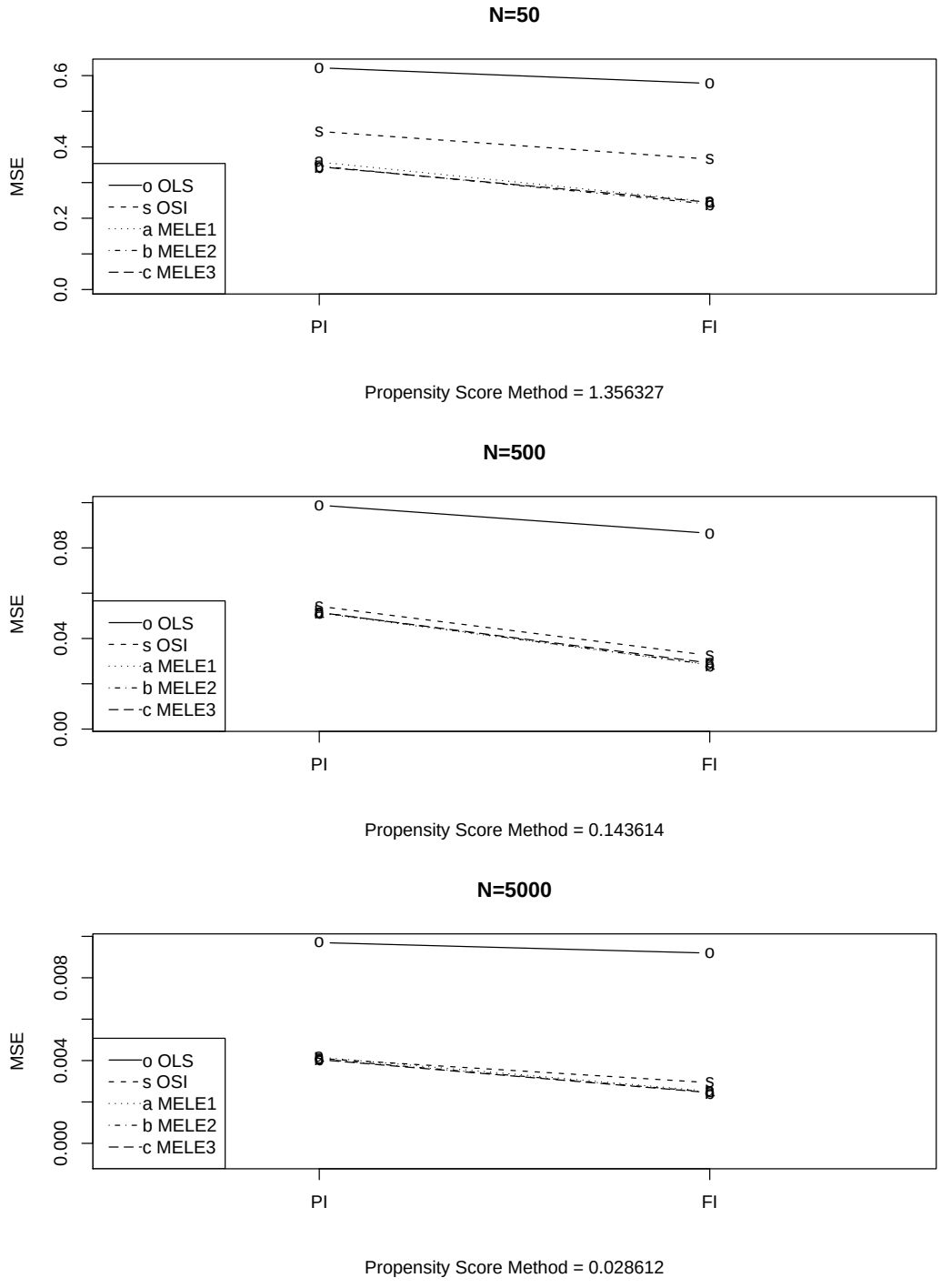

Fig. 22. MSE for estimating $E[Y]$ where the errors have the t distribution and an exponential missing structure

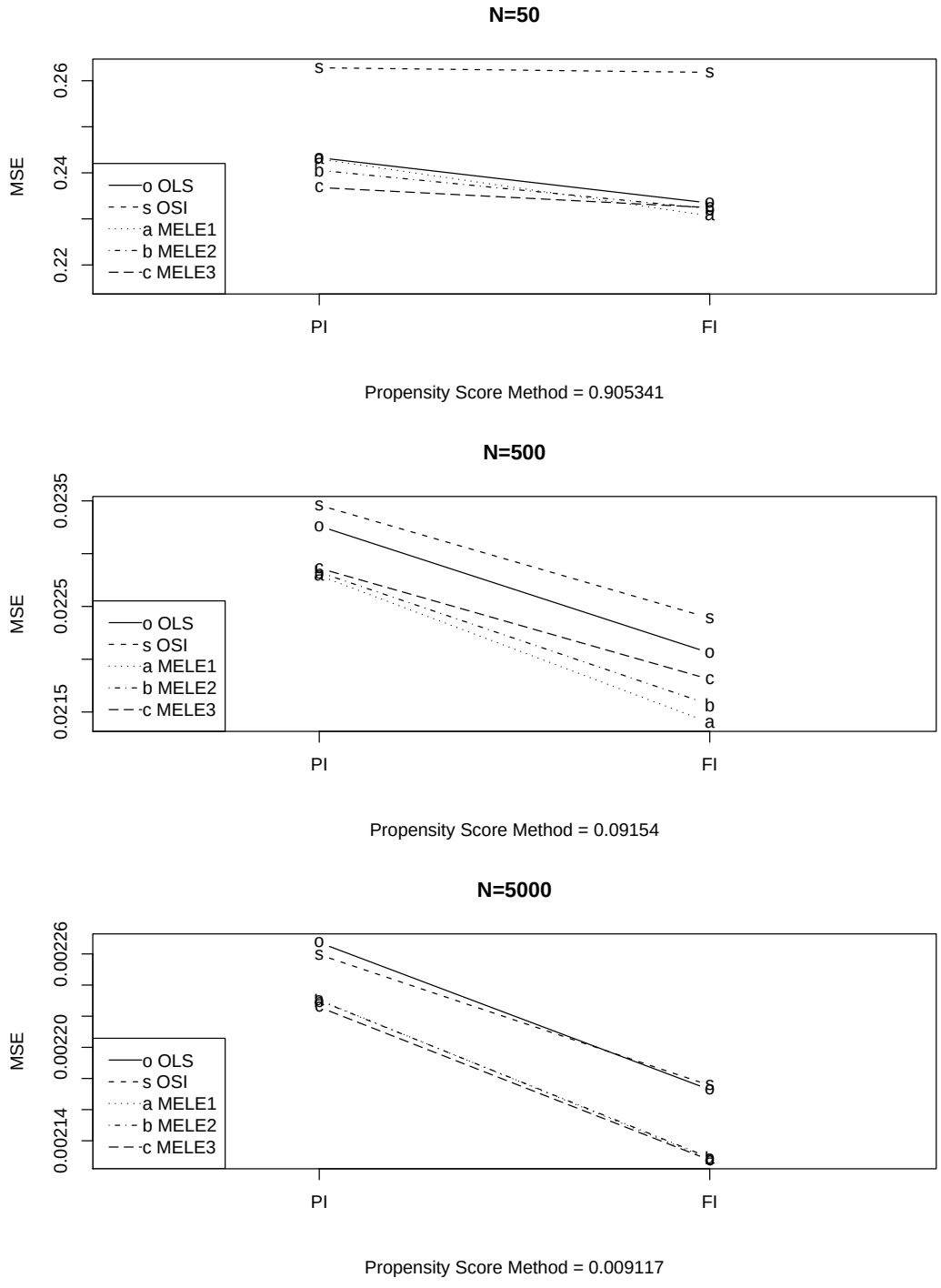

Fig. 23. MSE for estimating $E[Y]$ where the errors have the logistic distribution and an exponential missing structure

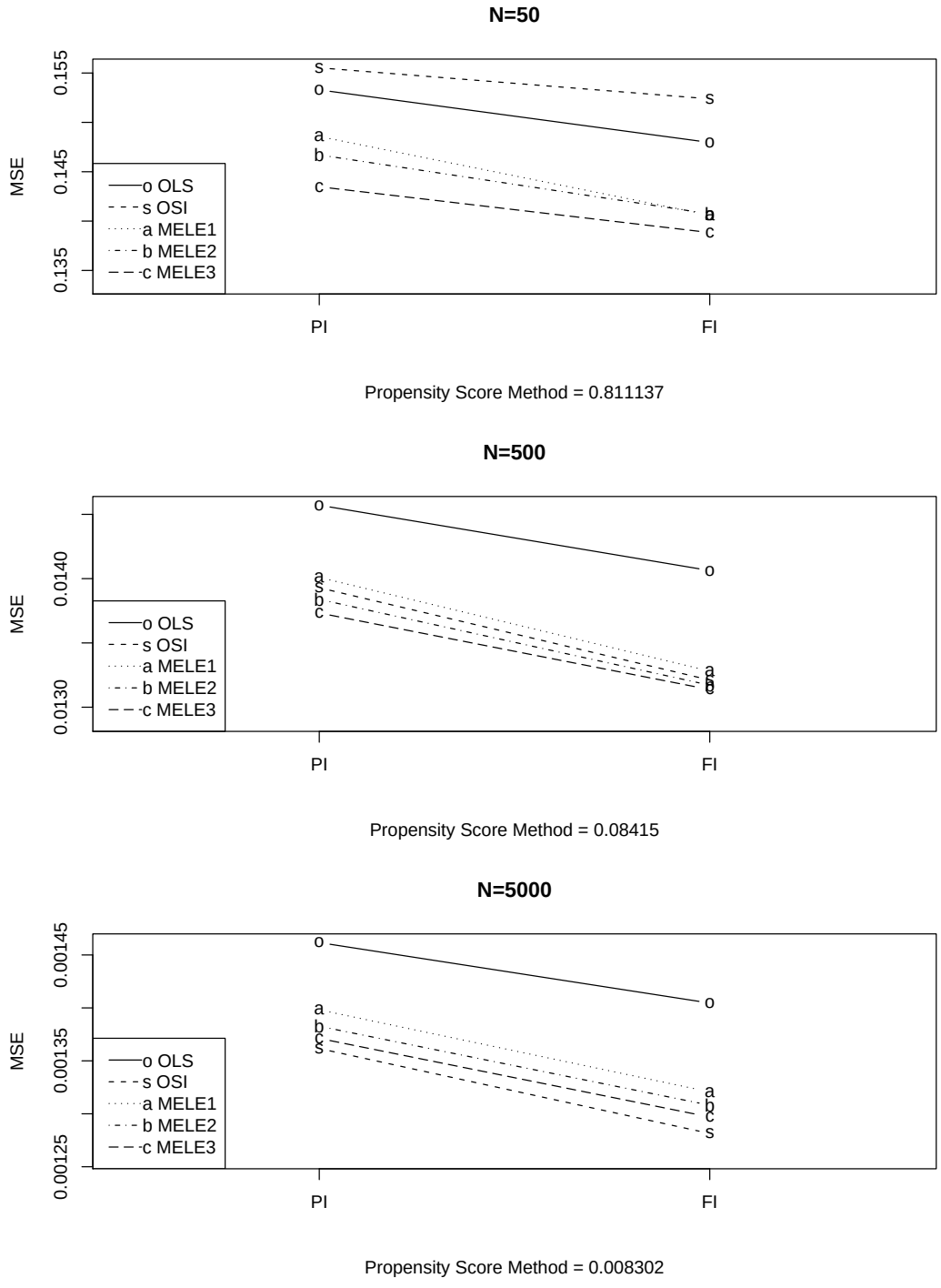

Fig. 24. MSE for estimating $E[Y]$ where the errors have the Gumbel distribution and an exponential missing structure

Table XVI. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is exponential and the errors have a normal distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.7173	0.1135	0.1102
	OSI	0.7173	0.1201	0.119
	MELE1	0.7173	0.1145	0.1107
	MELE2	0.7173	0.1145	0.1123
	MELE3	0.7173	0.1128	0.1115
N=500	OLS	0.5884	0.01117	0.01081
	OSI	0.5884	0.01138	0.01106
	MELE1	0.5884	0.01129	0.01093
	MELE2	0.5884	0.0113	0.01098
	MELE3	0.5884	0.01124	0.01095
N=5000	OLS	0.5772	0.001118	0.001089
	OSI	0.5772	0.001126	0.0011
	MELE1	0.5772	0.001128	0.0011
	MELE2	0.5772	0.001128	0.0011
	MELE3	0.5772	0.001122	0.001096

Table XVII. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is exponential and the errors have a t_2 distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	1.115	0.6219	0.5781
	OSI	1.115	0.4435	0.3657
	MELE1	1.115	0.3576	0.2465
	MELE2	1.115	0.3462	0.2396
	MELE3	1.115	0.3444	0.2448
N=500	OLS	0.6689	0.09887	0.08646
	OSI	0.6689	0.05422	0.0325
	MELE1	0.6689	0.05122	0.02831
	MELE2	0.6689	0.05156	0.02843
	MELE3	0.6689	0.05148	0.02917
N=5000	OLS	0.5844	0.009699	0.00919
	OSI	0.5844	0.0041	0.002932
	MELE1	0.5844	0.004163	0.002518
	MELE2	0.5844	0.004074	0.002474
	MELE3	0.5844	0.004043	0.002449

Table XVIII. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is exponential and the errors have a logistic distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.8189	0.2433	0.2334
	OSI	0.8189	0.2628	0.2618
	MELE1	0.8189	0.243	0.2307
	MELE2	0.8189	0.2405	0.2322
	MELE3	0.8189	0.2368	0.2325
N=500	OLS	0.6	0.02326	0.02206
	OSI	0.6	0.02346	0.02239
	MELE1	0.6	0.02279	0.0214
	MELE2	0.6	0.02282	0.02157
	MELE3	0.6	0.02286	0.02181
N=5000	OLS	0.5783	0.002267	0.002173
	OSI	0.5783	0.00226	0.002176
	MELE1	0.5783	0.00223	0.002128
	MELE2	0.5783	0.00223	0.002129
	MELE3	0.5783	0.002226	0.002128

Table XIX. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is exponential and the errors have a Gumbel distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.7453	0.1533	0.148
	OSI	0.7453	0.1555	0.1524
	MELE1	0.7453	0.1486	0.1406
	MELE2	0.7453	0.1467	0.1407
	MELE3	0.7453	0.1435	0.1388
N=500	OLS	0.594	0.01457	0.01406
	OSI	0.594	0.01393	0.01321
	MELE1	0.594	0.01401	0.01328
	MELE2	0.594	0.01384	0.01317
	MELE3	0.594	0.01373	0.01314
N=5000	OLS	0.5778	0.001462	0.001405
	OSI	0.5778	0.001362	0.001281
	MELE1	0.5778	0.001398	0.001321
	MELE2	0.5778	0.001383	0.001308
	MELE3	0.5778	0.001371	0.001297

Table XX. Simulation results showing the MSE for the estimation of $E[Y]$where the missing structure is exponential and the errors have a gamma distribution

		No Imputation	Partial Imputation	Full Imputation
N=50	OLS	0.7589	0.1702	0.1639
	OSI	0.7589	0.1629	0.1558
	MELE1	0.7589	0.159	0.1478
	MELE2	0.7589	0.154	0.1442
	MELE3	0.7589	0.1473	0.1377
N=500	OLS	0.5911	0.01654	0.01588
	OSI	0.5911	0.01423	0.01272
	MELE1	0.5911	0.015	0.01375
	MELE2	0.5911	0.01459	0.01343
	MELE3	0.5911	0.01418	0.013
N=5000	OLS	0.5776	0.001632	0.001567
	OSI	0.5776	0.001299	0.001136
	MELE1	0.5776	0.00141	0.001262
	MELE2	0.5776	0.001399	0.001269
	MELE3	0.5776	0.001372	0.001244

B. R code

1. Simulations with calculation of ϑ and E(Y)

The variable $filenum$ changes which sample size, missingness structure and error distribution is used. The final result creates a different file for each $filenum$.

```
for(filenum in 1:45){


library(evd)
library(nlme)
library(SparseM)
library(quantreg)
library(emplik)
gamma<--digamma(1)


###############DEFINE THE MODEL#########################
simulations=2000
nb<-1000
filewrite<-""
if(filenum<10){
filewritename<-paste("results//result0",filenum,".dat",sep="")
}else{
filewritename<-paste("results//result",filenum,".dat",sep="")
}
filewrite<-file(filewritename,open="w")
cat("Simulations:",simulations,file=filewrite,sep="")
cat("\n",file=filewrite)
```

```
nlist<-c(50,500,5000)
rlist<-c("tx")
elist<-c("normal","t2","logistic","gumbell","sgamma")
mlist<-c(1,2,3)
dlist<-c("e^zscore","1/(1+e^-x)","no missing")

n<-nlist[(filenum-1) %/% (length(elist)*length(dlist)) +1]
thisr<-rlist[1]
thise<-elist[(filenum-1) %/% (length(dlist)) %% length(elist) +1]
thisd<-dlist[(filenum-1) %% length(dlist) +1]

minx<-0
maxx<-2
expx<-(minx+maxx)/2
theta<-3

cat("y=",thisr,"+e e=",thise," N=",n,file=filewrite,sep="")
cat("\n",file=filewrite)
cat("t=",theta,", E[e]=0, E[x]=",expx,file=filewrite,sep="")
cat("\n",file=filewrite)
cat("missingness:",thisd,file=filewrite,sep="")
cat("\n","\n",file=filewrite)

##Define theta, r (parametric pieces)
##CHANGE PARAMETER ESTIMATION IF YOU CHANGE THE MODEL FORM
```

```
if(thisr=="tx"){
r<-function(t,x){
t<-as.numeric(t)
x<-as.numeric(x)
t*x
}
rdot<-function(t,x){
t<-as.numeric(t)
x<-as.numeric(x)
}
}
#Define h
h<-function(x,y){y}
trueh<-theta*expx

################################################

############ESTIMATE THE PARAMETER##################
othetas<-NULL
wthetas<-NULL
ethetas<-NULL
errorcatch<-0;
mthetas<-matrix(0,nrow=length(mlist),ncol=simulations)
thspartial<-NULL
ohspartial<-NULL
whspartial<-NULL
```

```
ehspartial<-NULL
mhspartial<-matrix(0,nrow=length(mlist),ncol=simulations)
hsno<-NULL
thsfull<-NULL
ohsfull<-NULL
whsfull<-NULL
ehsfull<-NULL
mhsfull<-matrix(0,nrow=3,ncol=simulations)
############START INDIVIDUAL SIMULATION###############
t<-0
while(t<simulations){
t<-t+1

##Define x (parametric covariate)
x<-runif(n,minx,maxx)

##Define e (error)
if(thise=="normal"){
e<-rnorm(n,0,1)
}
if(thise=="t2"){
e<-rt(n,2)
}
if(thise=="logistic"){
e<-rlogis(n,0,1)
}
```

```
if(thise=="gumbell"){
e<-rgumbel(n,0,1)-gamma
}
if(thise=="sgamma"){
e<-rgamma(n,2,1)-2
}

##Define y
y<-r(theta,x)+e

##Define the missingness
#dlist<-c("e^zscore","1/(1+e^-x)","no missing")
xscaled<-(x-mean(x))/sd(x)
if(thisd=="e^zscore"){
prob<--1/sqrt(2*pi)*exp(-xscaled^2)/dnorm(0)+1
prob[prob>.9]<-.9
prob[prob<.05]<-.05
plot(sort(x),prob[order(x)],pch=".",
ylim=c(0,1),type="l",xlab="x",ylab="e^(-x^2)")
pdf("testpdf.pdf")
}
if(thisd=="1/(1+e^-x)"){
prob<-1/(1+exp(-xscaled))
plot(sort(x),prob[order(x)],pch=".",ylim=c(0,1),type="l",
xlab="x",ylab="1/(1+e^-x)")
}
```

```
if(thisd=="no missing"){
prob<-rep(0,length(x))
}
d<-rbinom(n,1,1-prob)
#plot(sort(x),prob[order(x)],pch=".",ylim=c(0,1),
#main="Probability of missing",type="l")
#CHECK THAT IT'S NOT AN EMPTY DATASET
if(sum(d)<1){
d[2]<-1
}
if(sum(d)<2){
d[1]<-1
}
#plot(x[d==1],y[d==1],pch=1,
#xlim=c(min(x),max(x)),ylim=c(min(y),max(y)))
#rug(x[d==1],col="red")
#points(x[d==0],y[d==0],col="red",pch=19)



#TRUE
#theta<-theta

#OLSE
othetas[t]<-lm(y[d==1]~0+x[d==1])$coefficients[1]

#WLSE
```

```
glsy<-y[d==1]
glsx<-x[d==1]
test<-try(gls(glsy~0+glsx,correlation=corAR1())$coefficients[1],
silent=TRUE)
if(is.finite(test)){
wthetas[t]<-test
}else{
wthetas[t]<-NA
}

#ONE STEP IMPROVEMENT
#for y=x*Theta+e, score=-f'(e)/f(e)
rhat<-r(othetas[t],x)
rdothat<-rdot(othetas[t],x)
errors<-(y-rhat)
dens<-density(errors[d==1],n=1000)
index<-apply(abs(matrix(dens$x,nrow=length(dens$x),
ncol=length(errors))-
          matrix(errors,ncol=length(errors),
          nrow=length(dens$x),byrow=TRUE)),
2,which.min)
uindex<-index+1
lindex<-index-1
uindex[uindex>length(dens$x)]<-length(dens$x)-1
lindex[lindex<1]<-2
fe<-dens$y[index]
```

```
fep<-(dens$y[uindex]-dens$y[lindex])/
          (dens$x[uindex]-dens$x[lindex])
fep[fep=="NaN"]<-0
score<-(-1)*fep/fe
score[fe==0]<-0
#plot(errors,score)
sig2<-sum(d*errors^2)/sum(d)
muhat<-sum(d*rdothat)/sum(d)
xi<-(rdothat-muhat)*score+muhat*errors/var(errors)
ethetas[t]<-othetas[t]+sum(d*xi)/sum(d*xi^2)

#MELE
fbi<-NULL
nn<-sum(d==1)
#nb defined at top
bmin<-min(theta,ethetas[t])-abs(ethetas[t]-theta)*.5
bmax<-max(theta,ethetas[t])+abs(ethetas[t]-theta)*.5
testb<-seq(bmin,bmax,length=nb)
likes<-matrix(1000,length(mlist),nb)
for(b in 1:nb){
zi<-y[d==1]-testb[b]*x[d==1]
fbi<-NULL
for(ni in 1:nn)
fbi<-c(fbi,1/nn*sum(zi[ni]<=zi))
for(m in 1:length(mlist)){
phik<-sqrt(2)*cos(rep(seq(1,m),each=nn)*pi*fbi)
```

```
cikb<-matrix(c(zi,phik*(x[d==1]-mean(x[d==1]))),nrow=nn)
muvec<-rep(0,mlist[m]+1)
likes[m,b]<-as.numeric(el.test(cikb,muvec)[1])
}
}
minlike<-apply(likes,1,which.min)
mthetas[,t]<-testb[minlike]



##############################################
####ESTIMATE THE FUNCTION H ######################
##############################################

#THIS IS PARTIAL IMPUTATION
pimpute<-function(thetai){
thisy<-y
thisy[d==0]<-r(thetai,x[d==0])
mean(h(x,thisy))
}
thspartial[t]<-pimpute(theta)
ohspartial[t]<-pimpute(othetas[t])
whspartial[t]<-pimpute(wthetas[t])
ehspartial[t]<-pimpute(ethetas[t])
for(m in 1:length(mlist)){
mhspartial[t]<-pimpute(mthetas[m,t])
}
```

```
#THIS IS NO IMPUTATION
hsno[t]<-mean(h(x[d==1],y[d==1]))

#THIS IS FULL IMPUTATION
fullimputation<-function(thetai){
sumnum<-0
resid<-y-r(thetai,x)
thisf<-function(lambda){
sum(d*resid/(1 + (lambda*d*resid)))
}
bnds<-c((1/n-1)/max(resid[d==1]),(1/n-1)/min(resid[d==1]))
#yy<-NULL;xx<-seq(min(bnds)*2,max(bnds)*2,
#length=1000);for(xxx in xx){yy<-c(yy,thisf(xxx))};
#plot(xx,yy,ylim=c(-100,100),type="l");
#abline(v=min(bnds),col="green");abline(v=max(bnds),col="green");
#abline(h=0,col="red")
tolerance=10^(-9)
#thisu<-uniroot(thisf,bnds,tol=tolerance)
test<-try(uniroot(thisf,bnds,tol=tolerance)$root,silent=TRUE)
if(!is.finite(test)){
sumnum<-NA
errorcatch<-errorcatch+1
}else{
weights<-1/n*1/(1+(test*d*resid))
sumden<-sum(d)
```

```
for(i in 1:n){
sumnum<-sumnum
+sum(weights*d*h(x[i],r(thetai,x[i])+resid))
}
}
sumnum/sumden
}
thsfull[t]<-fullimputation(theta)
ohsfull[t]<-fullimputation(othetas[t])
whsfull[t]<-fullimputation(wthetas[t])
ehsfull[t]<-fullimputation(ethetas[t])
for(m in 1:length(mlist)){
mhsfull[m,t]<-fullimputation(mthetas[m,t])
}
}

#plot(density(hs),main=paste("Estimating h(x,y)=y"),
nodiff<-hsno-trueh
tpartdiff<-thspartial-trueh
tfulldiff<-thsfull-trueh
opartdiff<-ohspartial-trueh
ofulldiff<-ohsfull-trueh
wpartdiff<-whspartial-trueh
wfulldiff<-whsfull-trueh
epartdiff<-ehspartial-trueh
efulldiff<-ehsfull-trueh
```

```
mpartdiff<-matrix(0,nrow=length(mlist),ncol=simulations)
mfulldiff<-matrix(0,nrow=length(mlist),ncol=simulations)
for(m in 1: length(mlist)){
mpartdiff[m,]<-mhspartial[m,]-trueh
mfulldiff[m,]<-mhsfull[m,]-trueh
}

roundit<-12
msetheta<-as.data.frame(matrix(0,ncol=1,nrow=(3+length(mlist))))
colnames(msetheta)<-c("MSE")
rownames(msetheta)[1:3]<-c("WLS","OLS","OSI")
for(m in 1:length(mlist)){rownames(msetheta)[3+m]<-
paste("MELE",mlist[m],sep="")}
msetheta[1,]<-round(c(mean((wthetas-theta)^2)),roundit)
msetheta[2,]<-round(c(mean((othetas-theta)^2)),roundit)
msetheta[3,]<-round(c(mean((ethetas-theta)^2)),roundit)
for(m in 1:length(mlist)){
msetheta[3+m,]<-round(c(mean((mthetas[m,]-theta)^2)),roundit)
}

mseh<-as.data.frame(matrix(0,ncol=3,nrow=(4+length(mlist))))
colnames(mseh)<-c("None","Partial","Full")
rownames(mseh)[1:3]<-c("WLS","OLS","OSI")
for(m in 1:length(mlist)){rownames(mseh)[m+3]<-
paste("MELE",mlist[m],sep="")}
rownames(mseh)[nrow(mseh)]="TRUE"
```

```
mseh[1,]<-round(c(mean(nodiff^2),
mean(wpartdiff^2),mean(wfulldiff^2)),roundit)
mseh[2,]<-round(c(mean(nodiff^2),
mean(opartdiff^2),mean(ofulldiff^2)),roundit)
mseh[3,]<-round(c(mean(nodiff^2),
mean(epartdiff^2),mean(efulldiff^2)),roundit)
for(m in 1:length(mlist)){
mseh[3+m,]<-round(c(mean(nodiff^2),
mean(mpartdiff[m,]^2),mean(mfulldiff[m,]^2)),roundit)
}
mseh[nrow(mseh),]<-round(c(mean(nodiff^2),
mean(tpartdiff^2),mean(tfulldiff^2)),roundit)

write.table(msetheta,file=filewrite)
cat("\n",file=filewrite)
write.table(mseh,file=filewrite)
cat("\n","\n",file=filewrite)

#} #end for(n in nlist)
#} #end for(thisr in rlist)
#} #end for(thise in elist)
#} #end for(thisd in dlist)

cat("errors:",errorcatch)

if(filewrite!=""){
```

```
close(filewrite)
}
}
```

2. Combine output files

R-code used to take the multiple output data files and combine them together to find the average across all simulations. Some simulations were lost due to errors, and those are dropped and accounted for. The final output is a matrix called "$MSEEY$" which has for each row a scenario (a different combination of error structure, missingness structure, and sample size) and for each column a different estimate for ϑ (WLS, OLS, etc).

```
letters<-c("i","j","k","l","m","n","o","p","q","r","s","t","a1","a2",
"a3","a4","b1","b2","b3","b4")
WLSi<-matrix(0,nrow=45,ncol=length(letters))
OLSi<-matrix(0,nrow=45,ncol=length(letters))
OSIi<-matrix(0,nrow=45,ncol=length(letters))
MELE1i<-matrix(0,nrow=45,ncol=length(letters))
MELE2i<-matrix(0,nrow=45,ncol=length(letters))
MELE3i<-matrix(0,nrow=45,ncol=length(letters))
WLS<-NULL
OLS<-NULL
OSI<-NULL
MELE1<-NULL
MELE2<-NULL
MELE3<-NULL
```

```
N<-NULL
E<-NULL
D<-NULL
D2<-NULL
#EACH ROW IS A SCENARIO,
#EACH COLUMN A THETHAHAT-IMPUTE METHOD
MSEEY<-matrix(0,3*3*5,3*7)
emptymseey<-matrix(0,3*3*5,3*7)

#Combining across each file for a simluation
SIMSi<-matrix(0,nrow=45,ncol=length(letters))
nlist<-c(50,500,5000)
elist<-c("normal","t2","logistic","gumbell","sgamma")
dlist<-c("e^zscore","1/(1+e^-x)","no missing")
dlist2<-c("both","oneend","none")
ilist<-c("none","partial","full")
thetas<-c("OLS","OSI","MELE1","MELE2","MELE3")
tcolors<-c("black","green","blue","orange","red")
names<-c("OLS","OSI","MELE1","MELE2","MELE3")
file1<-"http://www.stat.tamu.edu/~crawford/
efficient/Rcodeforgraphs/sim/"
findstring<-function(f,s){
found<-NULL
for(m in 1:nchar(as.character(s))){
if(identical(substr(s,m,m),f)){
found<-c(found,m)
```

```
}
}
found
}
#for each scenario
for(i in 1:45){
catchMSEEY<-list()
N[i]<-nlist[(i-1) %/% (length(elist)*length(dlist)) +1]
E[i]<-elist[(i-1)
%/% (length(dlist)) %% length(elist) +1]
D[i]<-dlist[(i-1) %% length(dlist) +1]
D2[i]<-dlist2[(i-1) %% length(dlist2) +1]
for(j in 1:length(letters)){
if(i<10){
filename<-paste(file1,"output",letters[j],"/results/
result0",i,".dat",sep="")
}else{
filename<-paste(file1,"output",letters[j],"/results/
result",i,".dat",sep="")
}
if(as.numeric(file.access(filename))==0){
test<-read.delim(filename)

if(!is.na(as.numeric(substr(test[11,1],7,100)))){
MSEEYi<-matrix(0,1,3*7)
SIMSi[i,j]<-as.numeric(substr(test[4,1],findstring(":",test[4,1])
```

```
+2,100))
WLSi[i,j]<-as.numeric(substr(test[6,1],5,100))
OLSi[i,j]<-as.numeric(substr(test[7,1],5,100))
OSIi[i,j]<-as.numeric(substr(test[8,1],5,100))
MELE1i[i,j]<-as.numeric(substr(test[9,1],7,100))
MELE2i[i,j]<-as.numeric(substr(test[10,1],7,100))
MELE3i[i,j]<-as.numeric(substr(test[11,1],7,100))
for(a in 1:7){
spaces<-findstring(" ",test[12+a,1])
MSEEYi[1,(a-1)*3+1]<-as.numeric(substr(test[12+a,1],spaces[1]
+1,spaces[2]-1))
MSEEYi[1,(a-1)*3+2]<-as.numeric(substr(test[12+a,1],spaces[2]
+1,spaces[3]-1))
MSEEYi[1,(a-1)*3+3]<-as.numeric(substr(test[12+a,1],spaces[3]
+1,100))
}
catchMSEEY<-c(catchMSEEY,list(MSEEYi))
}
}
}
WLS[i]<-sum(WLSi[i,]*SIMSi[i,])/sum(SIMSi[i,])
OLS[i]<-sum(OLSi[i,]*SIMSi[i,])/sum(SIMSi[i,])
OSI[i]<-sum(OSIi[i,]*SIMSi[i,])/sum(SIMSi[i,])
MELE1[i]<-sum(MELE1i[i,]*SIMSi[i,])/sum(SIMSi[i,])
MELE2[i]<-sum(MELE2i[i,]*SIMSi[i,])/sum(SIMSi[i,])
MELE3[i]<-sum(MELE3i[i,]*SIMSi[i,])/sum(SIMSi[i,])
```

```
sumnafree<-rep(0,21)
#CHECK FOR MISSING SCENARIOS
for(k in 1:length(catchMSEEY)){
for(b in 1:21){
if(!is.na(catchMSEEY[[k]][1,b])){
emptymseey[i,b]<-1
sumnafree[b]<-sumnafree[b]+SIMSi[i,k]
MSEEY[i,b]<-MSEEY[i,b]+catchMSEEY[[k]][1,b]*SIMSi[i,k]
}
}
}
MSEEY[i,]<-MSEEY[i,]/sum(SIMSi[i,])
}
MSEEY[emptymseey==0]<-NA
colnamelist<-NULL
for(i in 1:21){colnamelist[i]<-paste(rep(c("WLS",names,"TRUE"),
each=3)[i],rep(ilist,7)[i],sep="")}
colnames(MSEEY)<-colnamelist
rownamelist<-NULL
for(i in 1:45){rownamelist[i]<-paste(N[i],E[i],D2[i],sep="")}
rownames(MSEEY)<-rownamelist

#A FUNCTION TO MAKE SURE ALL THE SIMULATIONS
#HAVE BEEN FOUND UP TO AT LEAST 20000
grachecksims<-function(){
if(debug==0){
```

```
setwd("U://html//efficient//Rcodeforgraphs//sim")
file="checksims.dat"
write("",file)
for(i in 1:45){
printthis<-paste("SIM",i,":",sum(SIMSi[i,]),sep="")
write(printthis,file,append=TRUE,sep="")
}
}
if(debug==1){
printans<-NULL
for(i in 1:45){
printans<-rbind(printans,sum(SIMSi[i,]))
rownames(printans)[i]=paste("SIM",i,":",sep="")
}
printans
}
}
#checksims()
```

3. Graph the MSE for the estimation of ϑ

The "plotj" function does the plot for scenario j where each scenario is a unique sample size, missingness structure and error distribution combination.

```
#FUNCTION TO FIT A SMOOTH 1/MSE CURVE
fitoox<-function(x,y){
ox<-1/x
```

```
fit<-lm(y~0+ox)
fit$coefficient[1]
}

#FUNCTION TO MAKE THE PLOT
plotj<-function(j,xmin=0,xmax=200,ymin=0,ymax=1){
i<-(j*3+2*((j-1)%/%5)-2)%%16
eindex<-(i-1)%/%(length(dlist))%% length(elist)+1
dindex<-(i-1) %% length(dlist) +1
pdffilename<-paste("theta",elist[eindex],".pdf",sep="")
if(debug==0){
pdf(file=pdffilename)
}
scene<-paste(elist[eindex],"-",dlist[dindex],"\n")
cri<-(E==(elist[eindex]))&(D==dlist[dindex])
xx<-seq(0,max(nlist),length=10000);
#if(debug==0){
themain=""
#}else{
#themain=scene
#}
linestyles<-c(1,2,3,4,5)
plot(xx,fitoox(nlist,OLS[cri])*1/xx,
xlim=c(xmin,xmax),ylim=c(ymin,ymax),
xlab="n",ylab="MSE",
main=themain,type="l",lty=linestyles[5])
```

```
lines(xx,fitoox(nlist,OSI[cri])*1/xx,lty=linestyles[4])
lines(xx,fitoox(nlist,MELE1[cri])*1/xx,lty=linestyles[3])
lines(xx,fitoox(nlist,MELE2[cri])*1/xx,lty=linestyles[2])
lines(xx,fitoox(nlist,MELE3[cri])*1/xx,lty=linestyles[1])
#legend("topright",lty=1,col=tcolors,legend=thetas)
legend("topright",lty=(6-linestyles),legend=thetas)
if(debug==0){
dev.off()
}
}
if(debug==1){
par(mfrow=c(2,3))
}
if(debug==0){
par(mfrow=c(1,1))
setwd("U://html//efficient//Thesis//Pics")
}
par(mfrow=c(1,1))
#normal
plotj(11,5,50,.01,.12)
#t2 USED IN ARTICLE
plotj(12,0,60,0,1)
#logistic
plotj(13,7.5,15,.15,.3)
#gumbell
plotj(14,3.5,12,.08,.3)
```

```
#sgamma USED IN ARTICLE
plotj(15,3,12,.08,.4)
if(debug==1){
plot(1,1)
}
```

4. Create a table for the estimation of ϑ

Final result prints out the table in Latex code.

```
####MAKING THETA ESTIMATION TABLES
boxes<-"no missing"
row2s<-N
rows<-thetas
cols<-E
table<-NULL
for(box in unique(boxes)){
for(row2 in unique(row2s)){#MANUAL
#for(row in unique(rows)){
thisrow<-NULL
for(col in unique(cols)){
thisrow<-cbind(thisrow,rbind(
WLS[(box==D)&(row2==row2s)&(col==cols)],
OLS[(box==D)&(row2==row2s)&(col==cols)],
OSI[(box==D)&(row2==row2s)&(col==cols)],
MELE1[(box==D)&(row2==row2s)&(col==cols)],
MELE2[(box==D)&(row2==row2s)&(col==cols)],
```

```
MELE3[(box==D)&(row2==row2s)&(col==cols)]
))
}
#}
table<-rbind(table,thisrow)
}
}

getround<-function(x){
ans<-0
digits<-4
start<-0
if(!is.nan(x)&&!is.na(x)){
while(floor(x*(10^start))==0 && start<12){
start<-start+1
}
ans<-round(x,start+digits-1)
}else{
ans<-NaN
}
ans
}

for(row in 1:nrow(table)){
for(col in 1:ncol(table)){
table[row,col]<-getround(table[row,col])
```

```
}
}

if(debug==0){
setwd("U://html//efficient//Thesis//Thesispieces")
file="App-simtable.tex"
write("",file)
write("\\clearpage",file,append=TRUE)
write("\\appendix{Tables showing the results
from the simulation}",file,append=TRUE)
write("\\label{app:tab}",file,append=TRUE)
for(row in 1:nrow(table)){
if(row
%%(length(unique(rows))*length(unique(row2s)))==1){
write("\n\n",file,append=TRUE)
write("\\begin{table}",file,append=TRUE)
write("\\begin{center}",file,append=TRUE)
capinput<-""
capinput<-paste(capinput,
"\\caption{Simulation results for the estimation of
$\\param$",sep="")
#if(row
%/%(length(unique(rows))*length(unique(row2s)))==0){
#capinput<-paste(capinput,
" where the missing structure is on both ends",sep="")
#capinput<-paste(capinput,
```

```
"\\label{tab:thetaboth}",set="")
#}
#if(row
%%(length(unique(rows))*length(unique(row2s)))==1){
#capinput<-paste(capinput,
" where the missing structure is on one end",sep="")
#capinput<-paste(capinput,
"\\label{tab:thetaone}",set="")
#}
#if(row
%/%(length(unique(rows))*length(unique(row2s)))==2){
#capinput<-paste(capinput,
" where there is no missing data",sep="")
capinput<-paste(capinput,
"\\label{tab:thetano}",set="")
#}
capinput<-paste(capinput,"}",sep="")
write(capinput,file,append=TRUE)
write("\\begin{tabular}{|c|c||c|c|c|c|c|}",file,append=TRUE)
write("\\hline",file,append=TRUE)
write("&&Normal&$t_2$&Logistic&Gumbel&Gamma\\\\",
file,append=TRUE)
#write("\\vspace{-20pt}",file,append=TRUE)
write("&&errors&errors&errors&errors&errors\\\\",
file,append=TRUE)
write("\\hline \\hline",file,append=TRUE)
```

```
}
input<-""
if(row%%length(unique(rows))==1){
input<-paste(input,"\\multicolumn{1}{|c|}{",sep="")
input<-paste(input,
"\\multirow{",length(unique(rows)),"}{*}{N=",sep="")
input<-paste(input,unique(row2s)[row])
%/%length(unique(rows))
%%length(unique(row2s))+1],"}",sep="")
input<-paste(input,"}&",sep="")
}else{
input<-paste(input,"&",sep="")
}
input<-paste(input,rows[(row-1)
%%6+1],"&",sep="")
input2<-paste(table[row,],collapse="&")
input<-paste(input,input2,sep="")
input<-paste(input,"\\\\",sep="")
write(input,file,append=TRUE)
if((row+1)
%%length(unique(rows))==1){
write(paste("\\hline",sep=""),file,append=TRUE)
}else{
write(paste("\\cline{2-",length(unique(rows))
+1,"}",sep=""),file,append=TRUE)
}
```

```
if((row+1)
%%(length(unique(rows))*length(unique(row2s)))==1){
write("\\end{tabular}",file,append=TRUE)
write("\\end{center}",file,append=TRUE)
write("\\end{table}",file,append=TRUE)
}
}
}
```

5. Solve for the MSE of the propensity score method

The simulation results did not return the propensity score method, so this code finds the needed values.

```
#SIMULATE NO IMPUTATION############
simulations=20000
nlist<-c(50,500,5000)
elist<-c("normal","t2","logistic","gumbell","sgamma")
dlist<-c("e^zscore","1/(1+e^-x)","no missing")
theta<-3
answer<-NULL
#install.packages("evd",lib="U://R")
#library(evd,lib.loc="U://R")
library(evd)
gamma<--digamma(1)
for(filenum in 1:45){
n<-nlist[(filenum-1)
```

```
%/% (length(elist)*length(dlist)) +1]
thise<-elist[(filenum-1)
%/% (length(dlist)) %% length(elist) +1]
thisd<-dlist[(filenum-1) %% length(dlist) +1]
scene<-paste("n:",n,thise,thisd)
thisanswer<-NULL
for(t in 1:20000){
     x<-runif(n,0,2)
     if(thise=="normal"){e<-rnorm(n,0,1)}
     if(thise=="t2"){e<-rt(n,2)}
     if(thise=="logistic"){e<-rlogis(n,0,1)}
     if(thise=="gumbell"){e<-rgumbel(n,0,1)-gamma}
     if(thise=="sgamma"){e<-rgamma(n,2,1)-2}
     y<-theta*x+e
     xscaled<-(x-mean(x))/sd(x)
     if(thisd=="e^zscore"){
     prob<--1/sqrt(2*pi)*exp(-xscaled^2)/dnorm(0)+1
           prob[prob>.9]<-.9
          prob[prob<.05]<-.05
     }
     if(thisd=="1/(1+e^-x)"){
          prob<-1/(1+exp(-xscaled))
     }
     if(thisd=="no missing"){
          prob<-rep(0,length(x))
     }
```

```
     d<-rbinom(n,1,1-prob)
     if(sum(d)<1){
          d[2]<-1
     }
     if(sum(d)<2){
          d[1]<-1
     }
     hsno<-1/n*sum(d*y/(1-prob))
     thisanswer[t]<-hsno-theta
}
answer<-rbind(answer,mean(thisanswer^2))
rownames(answer)[nrow(answer)]<-scene
}
noimputes<-answer
```

6. Graphs of the E(Y)

The function "graphey" takes in a variable *scene* which determines the combination of error distribution and missingness structure. The variable *pickn* determines the sample size to graph.

```
graphey<-function(scene,pickn){
if(debug==0){
setwd("U://html//efficient//Thesis//Pics")
pdffilename<-
paste("ey",substr(rownames(MSEEY)[scene],3,100),
".pdf",sep="")
```

```
pdf(file=pdffilename,width=9,height=4)
}
#par(mfrow=c(1,3),pin=c(2,2))
par(mfrow=c(1,1))
#colors<-c("green","blue","black","red","orange")
nonone<-sort(c((2:6)*3,(2:6)*3-1))
thesescenes<-scene+(c(1,2,3)-1)*15
#ylims=c(min(MSEEY[thesescenes,nonone]
#[!is.na(MSEEY[thesescenes,nonone])]),
#max(MSEEY[thesescenes,nonone]
#[!is.na(MSEEY[thesescenes,nonone])]))

#for(pickn in 1:3){
scenario<-scene+(pickn-1)*15
ylims=c(min(MSEEY[scenario,nonone]
[!is.na(MSEEY[scenario,nonone])]),
max(MSEEY[scenario,nonone]
[!is.na(MSEEY[scenario,nonone])]))
if(scene==1 && pickn==1){ylims[1]<-ylims[1]/1.02}
if(scene==2 && pickn==1){ylims[1]<-ylims[1]/1.03}
if(scene==4 && pickn==1){ylims[1]<-ylims[1]/-1.1}
if(scene==4 && pickn==2){ylims[1]<-ylims[1]/2.1}
if(scene==4 && pickn==3){ylims[1]<-ylims[1]/1.6}
if(scene==5 && pickn==1){ylims[1]<-ylims[1]/20.03}
if(scene==5 && pickn==2){ylims[1]<-ylims[1]/10.03}
if(scene==5 && pickn==3){ylims[1]<-ylims[1]/-3.03}
```

```
if(scene==7 && pickn==1){ylims[1]<-ylims[1]/1.04}
if(scene==8 && pickn==1){ylims[1]<-ylims[1]/1.07}
if(scene==10 && pickn==1){ylims[1]<-ylims[1]/1.02}
if(scene==11 && pickn==1){ylims[1]<-ylims[1]/1.04}
if(scene==11 && pickn==2){ylims[1]<-ylims[1]/1.02}
if(scene==11 && pickn==3){ylims[1]<-ylims[1]/1.02}
if(scene==13 && pickn==3){ylims[1]<-ylims[1]/1.02}
if(scene==14 && pickn==1){ylims[1]<-ylims[1]/1.08}
if(scene==14 && pickn==2){ylims[1]<-ylims[1]/1.06}
if(scene==14 && pickn==3){ylims[1]<-ylims[1]/1.13}
#noimputeis<-biganswer[scene,1]
mainline=""
#if(debug==1){
mainline<-paste("N=",N[scenario]
," D=",D2[scenario]," E=",E[scenario],sep="")
#}
#if(debug==0){
###TOGGLE HERE TO GIVE ONLY THE SAMPLE SIZE
cat(mainline)
mainline<-paste("N=",N[scenario],sep="")
#}
linesstyles<-c(1,2,3,4,5)
linepoints<-c("o","s","a","b","c")
plot(c(1,2),MSEEY[scenario,nonone[c(1,2)]],
ylab="MSE",xaxt="n",
xlab="",xlim=c(.5,2.5),type="b",
```

```
lty=linesstyles[1],
ylim=ylims,main=mainline,
pch=linepoints[1],
sub=paste("Propensity Score Method =",
round(noimputes[scenario],6))
)
axis(1,at=c(1,2),labels=c("PI","FI"))
legend("bottomleft",lty=linesstyles,merge=FALSE,
ncol=1,x.intersp=.05,legend=paste(linepoints,names))
for(i in 1:4){
lines(c(1,2),MSEEY[scenario,nonone[c(2*i+1,2*i+2)]],
type="b",lty=linesstyles[i+1],pch=linepoints[i+1])
}
#}
if(debug==0){
dev.off()
}
}
#NORMAL ONE END N=50 USED IN ARTICLE
graphey(2,1)
#NORMAL ONE END N=500 USED IN ARTICLE
graphey(2,2)
#NORMAL ONE END N=5000 USED IN ARTICLE
graphey(2,3)
#NORMAL NONE N=50 USED IN ARTICLE
graphey(3,1)
```

```
#NORMAL NONE N=500 USED IN ARTICLE
graphey(3,2)
#NORMAL NONE N=5000 USED IN ARTICLE
graphey(3,3)
#GAMMA ONE END N=50 USED IN ARTICLE
graphey(14,1)
#GAMMA ONE END N=500 USED IN ARTICLE
graphey(14,2)
#GAMMA ONE END N=5000 USED IN ARTICLE
graphey(14,3)
```

7. Tables of MSE values for estimating E(Y)

```
MSEEY2<-matrix(c(t(MSEEY)),ncol=3,byrow=TRUE)
colnames(MSEEY2)<-c("None","Partial","Full")
rownamelist<-NULL
rownamelist2<-NULL
for(i in 1:(3*3*5*7)){
rownamelist[i]<-paste(rep(N,each=7)[i],
rep(E,each=7)[i],rep(D2,each=7)[i],
rep(c("WLS",names,"TRUE"),45)[i],sep="")
rownamelist2[i]<-paste(rep(E,each=7)[i],
rep(D2,each=7)[i],sep="")
}
rownames(MSEEY2)<-rownamelist

for(row in 1:nrow(MSEEY2)){
```

```
for(col in 1:ncol(MSEEY2)){
MSEEY2[row,col]<-getround(MSEEY2[row,col])
}
}
row2s<-N
rows<-names
cols<-E

fixorder<-NULL
for(i in 1:3){ #MISSINGNESS STRUCTURES
for(j in 1:5){ #ERROR DISTRIBUTIONS
index<-(i)+(j-1)*3
fixorder<-c(fixorder,c(1:7,106:112,211:217)+7*(index-1))
#fixorder<-c(fixorder,(c(1,16,31)+i-1)+((j-1)*3))
}
}
MSEEY2<-MSEEY2[fixorder,]
rownamelist2<-rownamelist2[fixorder]
caps1<-NULL
caps1[1]<-"where the missing structure is gaussian"
caps1[2]<-"where the missing structure is exponential"
caps1[3]<-"where there is no missing data"
caps2<-NULL
caps2[1]<-" and the errors have a normal distribution"
caps2[2]<-" and the errors have a $t_2$ distribution"
caps2[3]<-" and the errors have a Logistic distribution"
```

```
caps2[4]<-" and the errors have a Gumbell distribution"
caps2[5]<-" and the errors have a Gamma distribution"
caps<-NULL
for(i in 1:15){caps[i]<-paste(rep(caps1,each=5)[i],
rep(caps2,3)[i],sep="")}

rows<-c("WLS","OLS","OSI","MELE1","MELE2","MELE3","TRUE")
input<-""
input2<-""
file="App-simtableey.tex"
write("",file)
#write("\\clearpage",file,append=TRUE)
#write("\\appendix{
Tables showing the results from the simulation for E[y]}",
file,append=TRUE)
#write("\\label{app:tab2}",file,append=TRUE)
for(row in 1:nrow(MSEEY2)){
if(row%%(length(unique(rows))*
length(unique(row2s)))==1){
write("\n\n",file,append=TRUE)
write("\\renewcommand{\\myf}{\\scriptsize}",
file,append=TRUE)
write("\\begin{table}[ht]",file,append=TRUE)
write("\\begin{center}",file,append=TRUE)
capinput<-""
capinput<-paste(capinput,
```

```
"\\caption{Simulation results showing the MSE for
the estimation of $E[\\Y]$",sep="")
capinput<-paste(capinput,
caps[row%/%(length(unique(rows))*
length(unique(row2s)))+1],sep="")
capinput<-paste(capinput,"\\label{tab:",
rownamelist2[row],"}",sep="")
capinput<-paste(capinput,"}",sep="")
write(capinput,file,append=TRUE)
write("\\begin{tabular}{|c|c||c|c|c|}",file,
append=TRUE)
write("\\hline",file,append=TRUE)
write("&&\\myf No&\\myf Partial&\\myf Full\\\\\",
file,append=TRUE)
#write("\\vspace{-20pt}",file,append=TRUE)
write("&&\\myf Imputation&\\myf Imputation
&\\myf Imputation\\\\\",file,append=TRUE)
write("\\hline \\hline",file,append=TRUE)
}
input<-""
if(row%%length(unique(rows))==1){
input<-paste(input,"\\multicolumn{1}{|c|}{",
sep="")
input<-paste(input,"\\multirow{",
length(unique(rows))-2,"}{*}{N=",sep="")
input<-paste(input,unique(row2s)[
```

```
row%/%length(unique(rows))%%length(unique(row2s))+1],
"}}",sep="")
#input<-paste(input,"}&\\myf ",sep="")
}else{
if(row%%7>1){
input<-paste(input,"&\\myf ",sep="")
}
}
if(row%%7>1){
input<-paste(input,rows[(row-1)%%7+1],
"&\\myf ",sep="")
input2<-paste(MSEEY2[row,],collapse="&\\myf ")
}

input<-paste(input,input2,sep="")
if(row%%7>1){
input<-paste(input,"\\\\",sep="")
}
write(input,file,append=TRUE)
input<-""
input2<-""
if((row+1)%%length(unique(rows))==1){
write(paste("\\hline",sep=""),file,append=TRUE)
}else{
if(row%%7>1){
write(paste("\\cline{2-5}",sep=""),file,
```

```
append=TRUE)
}
}
if((row+1)%%(length(unique(rows))*
length(unique(row2s)))==1){
write("\\end{tabular}",file,append=TRUE)
write("\\end{center}",file,append=TRUE)
write("\\end{table}",file,append=TRUE)
}
}
```

8. The asymptotic variances of estimators for $E(Y)$

The asymptotic variance for each example use different imputation methods as well as different estimates of ϑ and different error structures. The asymptotic variance for the weighted least squares estimates using the poor choice of weights is also calculated.

```
t<-3
ex<-1
exx<-4/3

missings<-c("symmetric","exponential","gaussian")
missingness<-missings[2]

wlschoices<-c("no","yes")
wls<-wlschoices[2]
```

```
if(missingness=="symmetric"){
#FOR SYMMETRIC MISSINGNESS
ed<-1/2
edx<-1/2
edxx<-71/120
eooedx<-25/8
exxoedx<-445/96
}
if(missingness=="exponential"){
#FOR EXPONENTIAL MISSINGNESS
ed<-1/2
edx<-0.37355217
edxx<-.413771
eooedx<-2.58059
exxoedx<-5.1455
}
if(missingness=="gaussian"){
#FOR GAUSSIAN MISSINGNESS
ed<-.5043435
edx<-.5043435
edxx<-.5801029
eooedx<-4.2222
exxoedx<-6.8661
}
if(wls=="yes"){
```

```
t<-3
dists<-c("U(0,2)","N(99,100)","Exp(1)","t3","Gamma(2,1)",
"Logistic","Gumbel")
ex<-c(1,99,1,0,2,0,.57721)
exx<-c(1.333,9901,1,1,6,0,1.978112)
}

gamma<--digamma(1)
getinfo<-function(errors){
     errors<-rnorm(10000)
     dens<-density(errors,n=1000)
     #INDEX MATCHES THE ERROR TO THE x DENSITY COORDINATE
     index<-apply(abs(matrix(dens$x,nrow=length(dens$x),
ncol=length(errors))-
      matrix(errors,ncol=length(errors),nrow=length(dens$x),
      byrow=TRUE)),2,which.min)
     h<-length(errors)/200
     uindex<-index+h
     uindex5<-index+2*h
     lindex<-index-h
     lindex5<-index-2*h
     uindex[uindex>length(dens$x)]<-lindex[uindex>length(dens$x)]
     lindex[lindex<1]<-uindex[lindex<1]
     uindex5[uindex5>length(dens$x)]<-
lindex5[uindex5>length(dens$x)]
     lindex5[lindex5<1]<-uindex5[lindex5<1]
```

```
     fe<-dens$y[index]
     fep<-(dens$y[uindex]-dens$y[lindex])/
     (dens$x[uindex]-dens$x[lindex])
     fep5<-(dens$y[lindex5]-8*dens$y[lindex]+
8*dens$y[uindex]-dens$y[uindex5])/
((dens$x[uindex]-dens$x[lindex])*6)
     fep[fep=="NaN"]<-0
     fep[uindex-lindex==0]<-0
     fep5[fep5=="NaN"]<-0
     fep5[uindex5-lindex5==0]<-0
     score<-(-1)*fep5/fe
     score[fep5==0]<-0
     #plot(errors,fep,pch=".")
     #points(errors,fep5,col="green",pch='.')
    #lines(seq(-4,4,length=300),
      #1/(sqrt(2*pi))*(-seq(-4,4,length=300))*
      #exp(-seq(-4,4,length=300)^2/2),col="red")
     mean(score^2)
}

getstats<-function(error){
library(evd,lib.loc="U://R")
library(evd)
variance<-0
information<-0
e<-NULL
```

```
nfore<-10000
if(error=="Uniform"){e<-runif(nfore,-1,1)}
if(error=="Normal"){e<-rnorm(nfore,0,1)}
if(error=="t3"){e<-rt(nfore,3)}
if(error=="Logistic"){e<-rlogis(nfore,0,1)}
if(error=="Gumbel"){e<-rgumbel(nfore,0,1)-gamma}
if(error=="Gamma"){e<-rgamma(nfore,2,1)-2}
if(error=="DExp"){e<-rexp(nfore,1)*(rbinom(nfore,1,.5)*2-1)}

c(var(e),getinfo(e))
}

timesten<-function(error){
ansv<-NULL
ansi<-NULL
pow<-10
for(i in 1:pow){
ans<-getstats(error)
ansv<-c(ansv,ans[1])
ansi<-c(ansi,ans[2])
}
c(mean(ansv),mean(ansi),sd(ansv),sd(ansi))}

dists<-c("Uniform","Normal","DExp","t3",
"Gamma","Logistic","Gumbel")
vs<-c(1/3,1,2,3,2,3.2899,1.6449)
```

```
is<-c(0,1,1,0.6667,1,0.3333,1)

getcheck<-function(){
check<-as.data.frame(matrix(0,8,length(dists)))
rownames(check)<-c("Theory v","Sim v","sd v","Diff v",
"Theory I","Sim I","sd I","Diff I")
colnames(check)<-dists
check[1,]<-round(vs,4)
check[5,]<-round(is,4)
for(i in 1:length(dists)){
getcheck<-round(timesten(dists[i]),4)
check[2,i]<-getcheck[1]
check[6,i]<-getcheck[2]
check[3,i]<-getcheck[3]
check[7,i]<-getcheck[4]}
check[4,]<-round(as.numeric(check[1,])-as.numeric(check[2,]),4)
check[8,]<-round(as.numeric(check[5,])-as.numeric(check[6,]),4)
check}

edzz<-function(v,i){edxx*i-edx*edx/ed*i+edx*edx/ed*1/v}
edzzs<-edzz(vs,is)

roundit<-3
avld<-function(v){
ans<-Inf
if(edx==ed*ex){
```

```
ans<-round((t*(edxx-edx*edx)*t+v*ed)/(ed*ed),roundit)
}ans}
avni<-function(v){
round(t*exxoedx*t+v*eooedx-t*ex*ex*t,roundit)}
avpiols<-function(v){round(t*(exx-ex*ex)*t+
v*ed+v*ex*edxx^(-1)*ex-v*edx*edxx^(-1)*edx,roundit)}
avpieff<-function(v,i){
round(t*(exx-ex*ex)*t+v*ed+ex*(edzz(v,i)^(-1))*ex
-edx*(edzz(v,i))^(-1)*edx,roundit)}
avfiols<-function(v){
round(t*(exx-ex*ex)*t+v*ex*(edxx)^(-1)*ex,roundit)}
avfieff<-function(v,i){
round(t*(exx-ex*ex)*t+ex*(edzz(v,i))^(-1)*ex,roundit)}
avpiwls<-function(v){round(t*(exx-ex*ex)*t+
v[2]*ed+v[2]*(ex-edx)*(exx-ex*ex)^(-1)*(ex-edx),roundit)}
avfiwls<-function(v){round(t*(exx-ex*ex)*t+
v[2]*ex*(exx-ex*ex)^(-1)*ex,roundit)}

ans<-as.data.frame(t(rbind(dists,avld(vs),avni(vs),
avpieff(vs,is),avfieff(vs,is),avpiols(vs),avfiols(vs))))
colnames(ans)<-c(missingness,"LD","NI",
"PIEFF","FIEFF","PIOLS","FIOLS")
ans
if(wls=="yes"){
ans<-as.data.frame(t(rbind(dists,avld(vs),
avni(vs),avpiwls(vs),avfiwls(vs))))
```

```
colnames(ans)<-c(paste("WLS:",missingness),
"LD","NI","PIWLS","FIWLS")
ans
}
```

C. Semi-parametric regression

1. Introduction

In this section I introduce a much more complex model, and go through the formulation for the canonical gradient. By matching this canonical gradient to the influence function of an estimator it can be used to check if an estimator is efficient for this model. Such an efficient estimator has yet to be found, but the following derivation shows the incredible complexity that comes with a generalization of the model. The model is

$$Y = r_\vartheta(X) + \gamma(Z) + \epsilon$$

where Y is the response, $r_\vartheta(X)$ is the known parametric function and ϑ is the unknown parameter for the random covariate X with dimension k_1, $\gamma(Z)$ is the unknown non-parametric component for the random covariate Z with dimension k_2, and ϵ is the random error term with distribution f which has mean 0 and variance σ^2. See Müller et al. (2007) for details on estimating the error distribution of such a model. Further assume X and Z are independent of ϵ.

The responses are MAR where

$$\delta = \begin{cases} 0 & \text{if } Y \text{ is missing} \\ 1 & \text{if } Y \text{ is not missing} \end{cases}$$

the observed data is $(X, Z, \delta, \delta Y)$. See Wang et al. (2004) for details on the Missing At Random assumption for this model. The goal will be to estimate $E[Y]$. The joint

distribution of the data is

$$P(X,Z,Y,\delta) = G(dx,dz)B_{\pi(X,Z)}(d\delta)\big\{\delta Q(dy|X,Z) + (1-\delta)\delta_o(Y)\big\}$$

where $G(dx,dz)$ is the marginal distribution of (X,Z), $B_{\pi(X,Z)}(d\delta)$ is the distribution of the conditional probability $\pi(X,Z)$ of $\delta = 1$, $Q(dy|X,Z)$ is the conditional distribution of Y given $(X = x, Z = z)$, and $\delta_o(Y)$ is the Dirac measure on dy at 0.

The canonical gradient is calculated by using Hellinger Derivatives to perturb the joint distribution and find the tangent space.

2. Perturbations through Hellinger derivatives

Lemma VI.1 *The tangent space defined for the model above is*

$$\dot{P}_n = \bigg\{u(X,Z) + \delta v(Y,X,Z) + (\delta - \pi)w(X,Z)\bigg\}.$$

where $u(X,Z)$, $v(Y,X,Z)$, and $w(X,Z)$ are perturbations defined as the following Hellinger Derivatives

$$\begin{aligned}
G_{nu}(X,Z) &= G(dx,dz)\big\{1 + n^{-1/2}u(X,Z)\big\}\\
Q_{nv}(Y|X,Z) &= Q(dy|X,Z)\big\{1 + n^{-1/2}v(Y,X,Z)\big\}\\
B_{\pi n\pi(X,Z)}(d\delta) &= B_{\pi(X,Z)}(d\delta)\big[1 + n^{-1/2}\big\{\delta - \pi(X,Z)\big\}w(X,Z)\big]
\end{aligned}$$

with the constraints that each perturbation is square integrable and

$$\int u(x,z)G(dx,dz) = 0 \tag{A.1}$$

$$\int v(y,x,z)Q(dy|x,z) = 0. \tag{A.2}$$

PROOF: The use of $\sqrt{n}$ consistency in a semi-parametric model is explained in Schick (1996a). Use the notation given above for the model and the derivatives with each constraint. It may be also noted that

$$Q(dy|X,Z) = f\left\{Y - r_\vartheta(X) - \gamma(Z)\right\} dy \tag{A.3}$$

which will be incorporated later. The proof will show that

$$P_{nuwv}(X,Z,Y,\delta) = P(X,Z,Y,\delta)(1 + n^{-1/2}\dot{P}_n)$$

The perturbed probability model is defined by

$$\begin{aligned}
& P_{nuwv}(X,Z,Y,\delta) \\
= \;& G_{nu}(X,Z)B_{\pi n\pi(X,Z)}(d\delta)\{\delta Q_{nv}(Y|X,Z) + (1-\delta)\delta_o(Y)\} \\
= \;& G(dx,dz)\{1 + n^{-1/2}u(X,Z)\}B_{\pi(X,Z)}(d\delta)[1 + n^{-1/2}\{\delta - \pi(X,Z)\}w(X,Z)] \\
& [\delta Q(dy|X,Z)\{1 + n^{-1/2}v(Y,X,Z)\} + (1-\delta)\delta_o(Y)].
\end{aligned}$$

This expands into

$$
\begin{aligned}
& P_{nuwv}(X,Z,Y,\delta) \\
= \; & G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)\delta n^{-1/2}v(Y,X,Z) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)(1-\delta)\delta_o(Y) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1/2}\{\delta-\pi(X,Z)\}w(X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1}\{\delta-\pi(X,Z)\}w(X,Z)\delta v(Y,X,Z) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)n^{-1/2}\{\delta-\pi(X,Z)\}w(X,Z)(1-\delta)\delta_o(Y) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1/2}u(X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1}u(X,Z)\delta v(Y,X,Z) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)n^{1/2}u(X,Z)(1-\delta)\delta_o(Y) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1}u(X,Z)\{\delta-\pi(X,Z)\}w(X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-3/2}u(X,Z)\{\delta-\pi(X,Z)\}w(X,Z)\delta v(Y,X,Z) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)n^{-1}u(X,Z)\{\delta-\pi(X,Z)\}w(X,Z)(1-\delta)\delta_o(Y).
\end{aligned}
$$

Any term with n to a power greater than $1/2$ can be put into $o_p(n^{-1/2})$, so

$$
\begin{aligned}
& P_{nuwv}(X,Z,Y,\delta) \\
= \; & G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)\delta n^{-1/2}v(Y,X,Z) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)(1-\delta)\delta_o(Y) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1/2}\{\delta-\pi(X,Z)\}w(X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)n^{-1/2}\{\delta-\pi(X,Z)\}w(X,Z)(1-\delta)\delta_o(Y) \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1/2}u(X,Z)\delta \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)n^{1/2}u(X,Z)(1-\delta)\delta_o(Y) \\
& +o_p(n^{-1/2}).
\end{aligned}
$$

After combining the like terms this becomes

$$
\begin{aligned}
& P_{nuwv}(X,Z,Y,\delta) \\
= \; & G(dx,dz)B_{\pi(X,Z)}(d\delta)\{\delta Q(dy|X,Z)+(1-\delta)\delta_o(Y)\} \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)Q(dy|X,Z)n^{-1/2} \\
& \left[\delta u(X,Z)+\delta v(Y,X,Z)+\{\delta-\pi(X,Z)\}w(X,Z)\delta\right] \\
& +G(dx,dz)B_{\pi(X,Z)}(d\delta)n^{-1/2}(1-\delta)\delta_o(Y)[\{\delta-\pi(X,Z)\}w(X,Z)+u(X,Z)] \\
& +o_p(n^{-1/2}).
\end{aligned}
$$

The first term is $P(X,Z,Y,\delta)$. Since $\delta(1-\delta)$ is 0 with probability 1, a term is

introduced at the end,

$$
\begin{aligned}
& P_{nuwv}(X,Z,Y,\delta) \\
= \; & P(X,Z,Y,\delta) \\
& +n^{-1/2}G(dx,dz)B_{\pi(X,Z)}(d\delta) \\
& \left(\delta Q(dy|X,Z)\Big[u(X,Z)+\delta v(Y,X,Z)+\{\delta-\pi(X,Z)\}w(X,Z)\Big]\right) \\
& +n^{-1/2}G(dx,dz)B_{\pi(X,Z)}(d\delta)\Bigg((1-\delta)\delta_o(Y)[u(X,Z)+ \\
& \{\delta-\pi(X,Z)\}w(X,Z)]\Bigg) \\
& +n^{-1/2}G(dx,dz)B_{\pi(X,Z)}(d\delta)\delta(1-\delta)\delta_o(Y)v(Y,X,Z) \\
= \; & P(X,Z,Y,\delta) \\
& +n^{-1/2}G(dx,dz)B_{\pi(X,Z)}(d\delta)\Big\{\delta Q(dy|X,Z)+(1-\delta)\delta_o(Y)\Big\} \\
& \Big[u(X,Z)+\delta v(Y,X,Z)+\{\delta-\pi(X,Z)\}w(X,Z)\Big] \\
& +o_p(n^{-1/2}) \\
= \; & P(X,Z,Y,\delta)\Big(1+n^{-1/2}\big[u(X,Z)+\delta v(Y,X,Z)+\{\delta-\pi(X,Z)\}w(X,Z)\big]\Big) \\
& +o_p(n^{-1/2}).
\end{aligned}
$$

Therefore the tangent space of $P(X,Z,Y,\delta)$ is

$$\dot{P}_n = \Big\{u(X,Z)+\delta v(Y,X,Z)+\{\delta-\pi(X,Z)\}w(X,Z)\Big\}$$

where

$$
\begin{aligned}
& u(X,Z) \in L_{2,0}(G) \\
& v(Y,X,Z) \in L_{2,0}(Q) \\
& w(X,Z) \in L_{2,0}(B).
\end{aligned}
$$

■

3. Hajek-Le Cam convolution theorem

An estimator $\hat{\chi}$ for $\chi(G, B, Q)$ is regular with limit $\mathcal{L}$ if $\mathcal{L}$ is a random variable such that under $P_{nuwv}(X, Z, Y, \delta)$

$$n^{1/2}\{\hat{\chi} - \chi(G, B, Q)\} \to \mathcal{L}.$$

The convolution theorem says that $\mathcal{L}$ is distributed as the sum of two random variables, the first being normal with mean zero and variance $E[(gr_{(\vartheta,\gamma)})^2]$, and the second being independent to the first. Thus $\hat{\chi}$ is efficient if it is regular with limit $N(0, E[(gr_{(\vartheta,\gamma)})^2])$. Using the definition of the influence function $\text{inf} \in L_{2,0}(P)$,

$$n^{-1/2}\{\hat{\chi} - \chi(G_{nu}, Q_{nw}, Q_{nv})\} \to n^{-1/2} \sum_{i=1}^{n} \text{inf} + o_p(1).$$

This means a regular estimator is efficient if and only if it is asymptotically linear with influence function $\text{inf} = gr_{(\vartheta,\gamma)}$ where $gr_{(\vartheta,\gamma)}$ is the canonical gradient function. The canonical gradient is defined by

$$n^{1/2}\{\hat{\chi} - \chi(G_{nu}, Q_{nw}, Q_{nv})\} \to E\{gr^{\top}_{(\vartheta,\gamma)} gr_{(\vartheta_*,\gamma_*)}\}. \tag{A.4}$$

where $gr_{(\vartheta,\gamma)}$ is any gradient in the the tangent space, and $gr_{(\vartheta_*,\gamma_*)}$ is the canonical gradient. The canonical gradient is a gradient which is in the tangent space of the model. The tangent space is found by perturbing the unknown elements of the model. For the model under consideration the gradient will be an element of $\dot{P}_n$, so define the canonical gradient as

$$gr_{(\vartheta_*,\gamma_*)} = u_*(X, Z) + \delta v_*(Y, X, Z) + \{\delta - \pi(X, Z)\} w_*(X, Z). \tag{A.5}$$

The following lemma will simplify the characterization used by Equation A.4 for this canonical gradient.

Lemma VI.2 *For the model defined above, the characterization for the canonical gradient is*

$$\begin{aligned} & E\{Yu(X,Z)\} + E\{Yv(Y,X,Z)\} \\ = \; & E\left\{u_*(X,Z)u(X,Z)\right\} + E\left\{\delta v_*(Y,X,Z)v(Y,X,Z)\right\} \\ & + E\left[\{\delta - \pi(X,Z)\}^2 w_*(X,Z)w(X,Z)\right]. \end{aligned}$$

Proof: For this model,

$$\begin{aligned} \chi(G_{nu}, Q_{nw}, Q_{nv}) &= E_{nuwv}[Y] &&= \int\int\int Y G_{nu}(X,Z) B_{\pi n \pi(X,Z)}(d\delta) Q_{nv}(Y|X,Z) \\ \chi(G,B,Q) &= E[Y] &&= \int\int\int Y G(dx,dz) B_{\pi(X,Z)}(d\delta) Q(dy|X,Z). \end{aligned}$$

This means the right hand side of Equation A.4 becomes

$$\begin{aligned} & \sqrt{n}\left\{\hat{\chi} - \chi(G,B,Q)\right\} \\ = \; & \sqrt{n}\Big\{\int\int\int Y G_{nu}(X,Z) B_{\pi n \pi(X,Z)}(d\delta) Q_{nv}(Y|X,Z) \\ & - \int\int\int Y G(dx,dz) B_{\pi(X,Z)}(d\delta) Q(dy|X,Z)\Big\}. \end{aligned}$$

Since Y does not depend on δ, $\pi(X,Z)$ or $w(X,Z)$,

$$\begin{aligned}
& \sqrt{n}\left\{\hat{\chi}-\chi(G,B,Q)\right\} \\
= \; & \sqrt{n}\Big\{\int\int YG_{nu}(X,Z)Q_{nv}(Y|X,Z)-\int\int YG(dx,dz)Q(dy|X,Z)\Big\} \\
= \; & \sqrt{n}\Big[\int\int YQ(dy|X,Z)\{1+n^{-1/2}v(Y,X,Z)\}G(dx,dz)\{1+n^{-1/2}u(X,Z)\} \\
& -\int\int YQ(dy|X,Z)G(dx,dz)\Big] \\
= \; & \sqrt{n}\Big[\int\int\Big\{YQ(dy|X,Z)G(dx,dz)+YQ(dy|X,Z)n^{-1/2}v(Y,X,Z)G(dx,dz) \\
& +YQ(dy|X,Z)G(dx,dz)n^{-1/2}u(X,Z)+ \\
& YQ(dy|X,Z)G(dx,dz)n^{-1}u(X,Z)v(Y,X,Z) \\
& -YQ(dy|X,Z)G(dx,dz)\Big\}\Big].
\end{aligned}$$

The term with n^{-1} will approach zero, so

$$\begin{aligned}
& \sqrt{n}\left\{\hat{\chi}-\chi(G,B,Q)\right\} \\
\to \; & \frac{\sqrt{n}}{\sqrt{n}}\int\int\Big[Y\{u(X,Z)+v(Y,X,Z)\}G(dx,dz)Q(dy|X,Z)\Big] \\
= \; & E\{Yu(X,Z)\}+E\{Yv(Y,X,Z)\}. \qquad \text{(A.6)}
\end{aligned}$$

The right hand side of Equation A.4 can be simplified by Lemma VI.1,

$$\begin{aligned}
E\left\{gr_{(\vartheta_*,\gamma_*)}gr_{(\vartheta,\gamma)}\right\} = \; & E\Big(\big[u_*(X,Z)+\delta v_*(Y,X,Z)+\{\delta-\pi(X,Z)\}w_*(X,Z)\big] \\
& \big[u(X,Z)+\delta v(Y,X,Z)+\{\delta-\pi(X,Z)\}w(X,Z)\big]\Big)
\end{aligned}$$

Since these perturbations were constrained to have a mean of zero, the cross product

terms will be zero. Therefore

$$\begin{aligned} & E\left\{gr_{(\vartheta_*,\gamma_*)}gr_{(\vartheta,\gamma)}\right\} \\ = \; & E\Big[u_*(X,Z)u(X,Z)+\delta v_*(Y,X,Z)v(Y,X,Z)+ \\ & \{\delta-\pi(X,Z)\}^2 w_*(X,Z)w(X,Z)\Big] \\ = \; & E\left\{u_*(X,Z)u(X,Z)\right\}+E\left\{\delta v_*(Y,X,Z)v(Y,X,Z)\right\} \\ & +E\left[\{\delta-\pi(X,Z)\}^2 w_*(X,Z)w(X,Z)\right]. \end{aligned} \tag{A.7}$$

Equating Equation A.6 to Equation A.7 we get

$$\begin{aligned} & E\{Yu(X,Z)\}+E\{Yv(Y,X,Z)\} \\ = \; & E\left\{u_*(X,Z)u(X,Z)\right\}+E\left\{\delta v_*(Y,X,Z)v(Y,X,Z)\right\} \\ & +E\left[\{\delta-\pi(X,Z)\}^2 w_*(X,Z)w(X,Z)\right]. \end{aligned}$$

■

Lemma VI.3 *For the canonical gradient defined in Equation A.5,* $w_*(X,Z)=0$*, so the canonical gradient is*

$$gr_{(\vartheta_*,\gamma_*)} = u_*(X,Z)+\delta v_*(Y,X,Z).$$

PROOF: The characterization given in Lemma VI.2 is true for any $u(X,Z)$ and $v(Y,X,Z)$ so set $u(X,Z)=0$ and $v(Y,X,Z)=0$. This yields

$$0 = E\left[\{\delta-\pi(X,Z)\}^2 w_*(X,Z)w(X,Z)\right].$$

This equation must hold true for any $w(X,Z)$, so $w_*(X,Z)=0$. ■

Lemma VI.4 *For the canonical gradient defined in Lemma VI.3,*

$$u_*(X,Z) = r_\vartheta(X) + \gamma(Z) - E\{r_\vartheta(X) + \gamma(Z)\}.$$

This leaves the canonical gradient as

$$gr_{(\vartheta_*,\gamma_*)} = r_\vartheta(X) + \gamma(Z) - E\{r_\vartheta(X) + \gamma(Z)\} + \delta v_*(Y,X,Z).$$

PROOF: The characterization given in Lemma VI.2 is true for any $v(Y,X,Z)$. From Lemma VI.3 we know $w_*(X,Z) = 0$, so set $v(Y,X,Z) = 0$. This leaves the characterization of the canonical gradient as

$$E\{Yu(X,Z)\} = E\{u_*(X,Z)u(X,Z)\}.$$

The obvious solution to this equation is $u_*(X,Z) = Y$, but the restriction from Equation A.1 was that the expected value needs to be zero. The next obvious step is $Y - E(Y)$ but this no longer solves the equation. The solution that satisfies the equation and constraint is $E\{Y|(X,Z)\} - E(Y)$. This solves the equation because

$$\begin{aligned}
E\{u_*(X,Z)u(X,Z)\} &= E\Bigg(\Big[E\{Y|(X,Z)\} - E(Y)\Big]u(X,Z)\Bigg) \\
&= E\Big[E\{Y|(X,Z)\}u(X,Z)\Big] - E\{E(Y)u(X,Z)\} \\
&= E\Bigg(E\Big[E\{Y|(X,Z)\}u(X,Z)|(X,Z)\Big]\Bigg) - E(Y)E\{u(X,Z)\} \\
&= E\Big[E\{Yu(X,Z)|(X,Z)\}\Big] - E(Y)E\{u(X,Z)\} \\
&= E\{Yu(X,Z)\}.
\end{aligned}$$

The last step comes from $E\{u(X,Z)\} = 0$. This satisfies Constraint A.1 because

$$\begin{aligned}\int u_*(X,Z)G(dx,dz) &= E\Big[E\{Y|(X,Z)\} - E(Y)\Big] \\ &= E\Big[E\{Y|(X,Z)\}\Big] - E(Y) \\ &= E(Y) - E(Y) \\ &= 0.\end{aligned}$$

This means

$$\begin{aligned}u_*(X,Z) &= E\{Y|(X,Z)\} - E(Y) \\ &= r_\vartheta(X) + \gamma(Z) - E\{r_\vartheta(X) + \gamma(Z)\}.\end{aligned}$$ ∎

The characterization for the canonical gradient given in Lemma VI.4 is true for any u so setting $u = 0$ the result is

$$E\{Yv(Y,X,Z)\} = E\{\delta v_*(Y,X,Z)v(Y,X,Z)\}. \tag{A.8}$$

This equation is not easily solved. This problem can be broken into smaller pieces by incorporating the restriction in Equation A.3

$$Q(dy|X,Z) = f\left\{Y - r_\vartheta(X) - \gamma(Z)\right\}dy \tag{A.9}$$

and defining further perturbations. Define the perturbations using the following Hellinger Derivatives,

$$\begin{aligned}f_{ns}(\epsilon) &= f(\epsilon)\left\{1 + n^{-1/2}s(\epsilon)\right\} \\ \vartheta_{nt} &= \vartheta + n^{-1/2}t \\ \gamma_{ng}(Z) &= \gamma(Z) + n^{-1/2}g(Z).\end{aligned}$$

for $t \in \mathbb{R}^{k_1}$, $g(Z)$ maps $\mathbb{R}^{k_2} \to \mathbb{R}$, $s(\epsilon) \in \mathbb{R}$ and subject to the constraints

$$\int s(\epsilon) f(\epsilon) d\epsilon = 0 \tag{A.10}$$

$$\int \epsilon s(\epsilon) f(\epsilon) d\epsilon = 0. \tag{A.11}$$

Which guarantees the needed assumptions for a valid error distribution, namely

$$\begin{aligned} E\{f_{ns}(\epsilon)\} &= 0 \\ \int f_{ns}(\epsilon)\, d\epsilon &= 1. \end{aligned}$$

Lemma VI.5 *Using the notation defined above,*

$$v_*(Y, X, Z) = s(\epsilon) + l(\epsilon)\{t^\top X + g(Z)\}.$$

PROOF: The perturbed f is

$$\begin{aligned} f_{ns}(\epsilon_{tg}) &= f_{ns}\{Y - r_{\vartheta_{nt}}(X) - \gamma_{ng}(Z)\} \\ &= f\{Y - r_{\vartheta_{nt}}(X) - \gamma_{ng}(Z)\} \\ &\quad \left[1 + n^{-1/2} s\{Y - r_{\vartheta_{nt}}(X) - \gamma_{ng}(Z)\}\right] \\ &= f\left[Y - \{r_{\vartheta_{nt}}(X) + \gamma(Z) + n^{-1/2} g(Z)\}\right] \\ &\quad \left(1 + n^{-1/2} s\left[Y - \{r_{\vartheta_{nt}}(X) + \gamma(Z) + n^{-1/2} g(Z)\}\right]\right). \end{aligned} \tag{A.12}$$

To simplify note that by using a Taylor Series expansion where $\dot{r}_\vartheta(X)$ is the derivative

of $r_\vartheta(X)$,

$$\begin{aligned} r_{\vartheta_{nt}}(X) &= r_\vartheta(X) + (\vartheta_{nt} - \vartheta)^\top \dot{r}_\vartheta(X) + o_p(n^{-1/2}) \\ &= r_\vartheta(X) + (\vartheta + n^{-1/2}t - \vartheta)^\top \dot{r}_\vartheta(X) + o_p(n^{-1/2}) \\ &= r_\vartheta(X) + n^{-1/2}t^\top \dot{r}_\vartheta(X) + o_p(n^{-1/2}). \end{aligned}$$

By substituting this into Equation A.12

$$\begin{aligned} f_{ns}(\epsilon_{tg}) &= f\left[Y - \{r_\vartheta(X) + n^{-1/2}t^\top \dot{r}_\vartheta(X) + \gamma(Z) + n^{-1/2}g(Z)\}\right] \\ &\quad \left(1 + n^{-1/2}s\left[Y - \{r_\vartheta(X) + n^{-1/2}t^\top \dot{r}_\vartheta(X) + \gamma(Z) + n^{-1/2}g(Z)\}\right]\right) \\ &= f\left[Y - \{r_\vartheta(X) + \gamma(Z) + n^{-1/2}t^\top \dot{r}_\vartheta(X) + n^{-1/2}g(Z)\}\right] \\ &\quad \left(1 + n^{-1/2}s\left[Y - \{r_\vartheta(X) + \gamma(Z) + n^{-1/2}t^\top \dot{r}_\vartheta(X) + n^{-1/2}g(Z)\}\right]\right). \end{aligned}$$

To find the relationship between $f_{ns}(\epsilon_{tg})$ and $f(\epsilon)$ I will use a Taylor Series expansion. To help simplify the calculation let $\Delta = -n^{-1/2}\{t^\top \dot{r}_\vartheta(X) + g(Z)\}$. Then equation becomes

$$f_{ns}(\epsilon_{tg}) = f(\epsilon + \Delta)\left\{1 + n^{-1/2}s(\epsilon + \Delta)\right\}. \tag{A.13}$$

Define $\dot{f}(\epsilon)$ as the derivative of $f(\epsilon)$ and $\dot{s}(\epsilon)$ as the derivative of $s(\epsilon)$. Then by Taylor Expansion

$$\begin{aligned} f(\epsilon + \Delta) &= f(\epsilon) + (\epsilon + \Delta - \epsilon)\dot{f}(\epsilon) + o_p(n^{-1/2}) \\ s(\epsilon + \Delta) &= s(\epsilon) + (\epsilon + \Delta - \epsilon)\dot{s}(\epsilon) + o_p(n^{-1/2}). \end{aligned}$$

Using these Taylor Expansions with Equation A.13 yields

$$\begin{aligned} f_{ns}\left(\epsilon_{tg}\right) &= \left\{f(\epsilon)+\Delta\dot{f}(\epsilon)\right\}\left[1+n^{-1/2}\left\{s(\epsilon)+\Delta\dot{s}(\epsilon)\right\}\right]+o_p(n^{-1/2}) \\ &= \left[f(\epsilon)-n^{-1/2}\{t^\top\dot{r}_\vartheta(X)+g(Z)\}\dot{f}(\epsilon)\right] \\ &\quad \left(1+n^{-1/2}\left[s(\epsilon)-n^{-1/2}\{t^\top\dot{r}_\vartheta(X)+g(Z)\}\dot{s}(\epsilon)\right]\right)+o_p(n^{-1/2}) \\ &= f(\epsilon)+f(\epsilon)n^{-1/2}s(\epsilon)-n^{-1}f(\epsilon)\dot{s}(\epsilon)\{t^\top\dot{r}_\vartheta(X)+g(Z)\} \\ &\quad -n^{-1/2}\dot{f}(\epsilon)\{t^\top\dot{r}_\vartheta(X)+g(Z)\}-\dot{f}(\epsilon)s(\epsilon)n^{-1}\{t^\top\dot{r}_\vartheta(X)+g(Z)\} \\ &\quad +n^{-3/2}\dot{f}(\epsilon)\dot{s}(\epsilon)\{t^\top\dot{r}_\vartheta(X)+g(Z)\}+o_p(n^{-1/2}). \end{aligned}$$

Note that any term with n raised to a negative power greater than $1/2$ can be put into the term $o_p(n^{-1/2})$. Then

$$\begin{aligned} f_{ns}\left(\epsilon_{tg}\right) &= f(\epsilon)+f(\epsilon)n^{-1/2}s(\epsilon)-n^{-1/2}\dot{f}(\epsilon)\{t^\top\dot{r}_\vartheta(X)+g(Z)\} \\ &= f(\epsilon)\left(1+n^{-1/2}\left[s(\epsilon)-\frac{\dot{f}(\epsilon)}{f(\epsilon)}\{t^\top\dot{r}_\vartheta(X)+g(Z)\}\right]\right). \end{aligned}$$

Define the score function $l(\epsilon)=-\frac{\dot{f}(\epsilon)}{f(\epsilon)}$, then

$$f_{ns}\left(\epsilon_{tg}\right) = f(\epsilon)\left(1+n^{-1/2}\left[s(\epsilon)+l(\epsilon)\left\{t^\top\dot{r}_\vartheta(X)+g(Z)\right\}\right]\right)+o_p(n^{-1/2})$$

By comparing this result with Equation A.2 and Equation A.3 and noting that this perturbation is a function of (X,Z,Y) it is clear

$$v_*(Y,X,Z)=s(\epsilon)+l(\epsilon)\left\{t^\top\dot{r}_\vartheta(X)+g(Z)\right\} \qquad \blacksquare$$

4. Simplifying the tangent space

Future calculations will be simplified by breaking $v_*(Y,X,Z)$ into two convenient pieces. To do this the following three lemmas will be helpful. The first shows that

any function of X and δ times ϵ is zero, the second shows $E\{\epsilon l(\epsilon)\} = 1$, and the third how to pull δ out of an expected value.

Lemma VI.6 *For the model as defined above for any function $g(X, \delta)$,*

$$E\left[g(X,\delta)\epsilon\right] = 0$$

PROOF:

$$\begin{aligned}
E\left[g(X,\delta)\epsilon\right] &= E\left[g(X,\delta)(Y - \vartheta^\top X)\right] \\
&= E\left[g(X,\delta)Y\right] - E\left[g(X,\delta)\vartheta^\top X\right] \\
&= E\left[E\left\{g(X,\delta)Y|X\right\}\right] - E\left[E\left\{g(X,\delta)\vartheta^\top X|X\right\}\right] \\
&= E\left[E\left\{g(X,\delta)|X\right\}E\left\{Y|X\right\}\right] - E\left[E\left\{g(X,\delta)|X\right\}\vartheta^\top X\right] \\
&= E\left[E\left\{g(X,\delta)|X\right\}\vartheta^\top X\right] - E\left[E\left\{g(X,\delta)|X\right\}\vartheta^\top X\right] \\
&= 0.
\end{aligned}$$

■

Lemma VI.7 *For any random variable ϵ with distribution $f(\epsilon)$, finite expected value, and score function $l(\epsilon)$,*

$$E\{\epsilon l(\epsilon)\} = 1.$$

PROOF:

$$\begin{aligned} E\{\epsilon l(\epsilon)\} &= \int \epsilon l(\epsilon) f(\epsilon) d\epsilon \\ &= -\int \epsilon \frac{\dot{f}(\epsilon)}{f(\epsilon)} f(\epsilon) d\epsilon \\ &= -\int \epsilon \dot{f}(\epsilon) d\epsilon. \end{aligned}$$

Then using integration by parts

$$\begin{aligned} E\{\epsilon l(\epsilon)\} &= -\Big\{\epsilon f(\epsilon)\Big\}_{-\infty}^{\infty} + \int f(\epsilon) d\epsilon \\ &= -\Big\{\epsilon f(\epsilon)\Big\}_{-\infty}^{\infty} + 1 \\ &= \lim_{\epsilon \to -\infty} \epsilon f(\epsilon) - \lim_{\epsilon \to \infty} \epsilon f(\epsilon) + 1 \end{aligned}$$

.

Since the expected value of ϵ is finite,

$$\int \epsilon f(\epsilon) d\epsilon = \int_M^{\infty} \epsilon f(\epsilon) d\epsilon + \int_{\infty}^{M} \epsilon f(\epsilon) d\epsilon < \infty$$

by the Divergence Theorem the limit of $\epsilon f(\epsilon)$ as ϵ goes to $-\infty$ or ∞ will be zero. Thus

$$E\{\epsilon l(\epsilon)\} = 1.$$ ■

Lemma VI.8 *For any function* $g(X, Z, \delta, Y)$

$$E\{\delta g(X, Z, \delta, Y)\} = E(\delta) E\{g(X, Z, \delta, Y) | \delta = 1\}.$$

PROOF: The proof is trivial since δ can only take two values, 0 or 1.

$$\begin{aligned} & E\{\delta g(X,Z,\delta,Y)\} \\ = \; & E\{1g(X,Z,\delta,Y)|\delta=1\}P(\delta=1) + E\{0g(X,Z,\delta,Y)\}P(\delta=1) \\ = \; & E\{g(X,Z,\delta,Y)|\delta=1\}E(\delta) \end{aligned}$$

■

With these two lemmas it will be easier to show the simplification of $v_*(Y,X,Z)$.

Lemma VI.9 *Using the notation above $v_*(Y,X,Z)$ can be rewritten as the sum of the following pieces*

$$s_2(\epsilon) + \xi(X,Z,\epsilon)$$

where

$$\begin{aligned} s_2(\epsilon) &\in \mathcal{S} = \{s(\epsilon)\} \\ \xi(X,Z,\epsilon) &= \Big[\phi(X,Z) - E\{\phi(X,Z)|\delta=1\}\Big] l(\epsilon) + E\{\phi(X,Z)|\delta=1\}\frac{\epsilon}{\sigma^2} \\ \phi(X,Z) &= t^\top \dot{r}_\vartheta(X) + g(Z) \end{aligned}$$

This is desirable because $s_2\epsilon$ is orthogonal to $\delta\xi(X,Z,\epsilon)$.

PROOF: Using the definitions of $\xi(X,Z,\epsilon)$ and $\phi(X,Z)$ given in the statement of the lemma, add the defintion

$$s_3(\epsilon) = E\{\phi(X,Z)|\delta=1\}\left\{l(\epsilon) - \frac{\epsilon}{\sigma^2}\right\}.$$

Note that $s_3(\epsilon) \in \int$ because it satisfies the constraints in Equation A.10 and Equation A.11, namely $E[s_3(\epsilon)] = 0$ and $E[\epsilon s_3(\epsilon)] = 0$. The form of $v_*(Y,X,Z)$ can be

rewritten as

$$\begin{aligned} v_*(Y,X,Z) &= s(\epsilon) + l(\epsilon)\left\{t^\top \dot{r}_\vartheta(X) + g(Z)\right\} \\ &= s(\epsilon) + E\{\phi(X,Z)|\delta=1\}\left\{l(\epsilon) - \frac{\epsilon}{\sigma^2}\right\} \\ &\quad + \Big[\phi(X,Z) - E\{\phi(X,Z)|\delta=1\}\Big] l(\epsilon) + E\{\phi(X,Z)|\delta=1\}\frac{\epsilon}{\sigma^2} \\ &= s(\epsilon) + s_3(\epsilon) + \xi(X,Z,\epsilon). \end{aligned}$$

Next I will show that $s(\epsilon) + s_3(\epsilon)$ belongs to the set of $s_3(\epsilon)$ by showing it satisfies the constraints in Equation A.10 and Equation A.11. First to show that $s(\epsilon) + s_3(\epsilon)$ satistifies Equation A.10,

$$\begin{aligned} \int \{s(\epsilon) + s_3(\epsilon)\} f(\epsilon) d\epsilon &= \int s(\epsilon) f(\epsilon) d\epsilon + \int s_3(\epsilon) f(\epsilon) d\epsilon \\ &= \int s_3(\epsilon) f(\epsilon) d\epsilon \\ &= \int E\{\phi(X,Z)|\delta=1\} l(\epsilon) f(\epsilon) d\epsilon \\ &\quad - \int E\{\phi(X,Z)|\delta=1\}\epsilon/\sigma^2 f(\epsilon) d\epsilon \\ &= E\Big[E\{\phi(X,Z)|\delta=1\} l(\epsilon)\Big] - \\ &\quad 1/\sigma^2 E\Big[E\{\phi(X,Z)|\delta=1\}\epsilon\Big] \\ &= E\Big(E\big[E\{\phi(X,Z)|\delta=1\}|(X,Z)\big] E\big\{l(\epsilon)|(X,Z)\big\}\Big) \\ &\quad -1/\sigma^2 E\Big(E\big[E\{\phi(X,Z)|\delta=1\}|(X,Z)\big] E\big\{\epsilon|(X,Z)\big\}\Big) \\ &= 0. \end{aligned}$$

And $s(\epsilon) + s_3(\epsilon)$ also satisfies Equation A.11 because

$$\begin{aligned}
\int \epsilon \left\{ s(\epsilon) + s_3(\epsilon) \right\} f(\epsilon) d\epsilon &= \int \epsilon s(\epsilon) f(\epsilon) d\epsilon + \int \epsilon s_3(\epsilon) f(\epsilon) d\epsilon \\
&= \int \epsilon s_3(\epsilon) f(\epsilon) d\epsilon \\
&= \int E\{\phi(X,Z)|\delta = 1\} \epsilon l(\epsilon) f(\epsilon) d\epsilon \\
&\quad - \int E\{\phi(X,Z)|\delta = 1\} \epsilon^2 / \sigma^2 f(\epsilon) d\epsilon \\
&= E\Big[E\{\phi(X,Z)|\delta = 1\} \epsilon l(\epsilon) \Big] \\
&\quad -1/\sigma^2 E\Big[E\{\phi(X,Z)|\delta = 1\} \epsilon^2 \Big] \\
&= E\Big(E\big[E\{\phi(X,Z)|\delta = 1\} | (X,Z) \big] E\big\{ \epsilon l(\epsilon) | (X,Z) \big\} \Big) \\
&\quad -1/\sigma^2 E\Big(E[E\{\phi(X,Z)|\delta = 1\}|(X,Z)] E\{\epsilon^2 | (X,Z)\} \Big)
\end{aligned}$$

Now using Lemma VI.7,

$$\begin{aligned}
\int \epsilon \left\{ s(\epsilon) + s_3(\epsilon) \right\} f(\epsilon) d\epsilon &= E\Big[E\{\phi(X,Z)|\delta = 1\} \Big] - \sigma^2/\sigma^2 E\Big[E\{\phi(X,Z)|\delta = 1\} \Big] \\
&= E\Big[E\{\phi(X,Z)|\delta = 1\} \Big] - E\Big[E\{\phi(X,Z)|\delta = 1\} \Big] \\
&= 0.
\end{aligned}$$

To simplify I will now refer to $s(\epsilon) + s_3(\epsilon)$ as simply $s_2(\epsilon)$ since it satisfies the conditions to be an element of the set $s(\epsilon)$. This simplifies Lemma VI.5 to $v_*(Y, X, Z) = s_2(\epsilon) + \xi(X, Z, \epsilon)$. This notation is desirable because the two pieces $s_2(\epsilon)$ and $\delta\xi(X, Z, \epsilon)$ are orthogonal. This can be shown by finding the expected value of the product, which

is

$$\begin{aligned}
& E\left\{\delta s_2(\epsilon)\xi(X,Z,\epsilon)\right\} \\
= \; & E\left\{\delta s_2(\epsilon)\left(\left[\phi(X,Z)-E\{\phi(X,Z)|\delta=1\}\right]l(\epsilon)+E\{\phi(X,Z)|\delta=1\}\frac{\epsilon}{\sigma^2}\right)\right\} \\
= \; & E\{\delta s_2(\epsilon)\phi(X,Z)l(\epsilon)\}-E\{\delta s_2(\epsilon)l(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\
& +\frac{1}{\sigma^2}E\{\delta s_2(\epsilon)\epsilon\}E\{\phi(X,Z)|\delta=1\} \\
= \; & E\{\delta\phi(X,Z)\}E\{s_2(\epsilon)l(\epsilon)\}-E(\delta)E\{s_2(\epsilon)l(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\
& +\frac{1}{\sigma^2}E(\delta)E\{s_2(\epsilon)\epsilon\}E\{\phi(X,Z)|\delta=1\}.
\end{aligned}$$

Using Equation A.11 the last term is zero, and Lemma VI.8 simplifies the first term, leaving

$$\begin{aligned}
& E\left\{\delta s_2(\epsilon)\xi(X,Z,\epsilon)\right\} \\
= \; & E\{\phi(X,Z)|\delta=1\}E(\delta)E\{s_2(\epsilon)l(\epsilon)\}-E(\delta)E\{s_2(\epsilon)l(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\
= \; & 0.
\end{aligned}$$

Thus the perturbation can be rewritten as the sum of

$$v_*(Y,X,Z)=s_2(\epsilon)+\xi(X,Z,\epsilon)$$

Where $s_2(\epsilon)$ is orthogonal to $\delta\xi(X,Z,\epsilon)$. ∎

Lemma VI.10 *The condition given in Equation A.8 can be simplified as*

$$E\{\phi(X,Z)\}=E[\delta\{s_{2*}(\epsilon)+\xi_*(X,Z,\epsilon)\}\{s_2(\epsilon)+\xi(X,Z,\epsilon)\}].$$

PROOF: Using Lemma VI.9 into Equation A.8 the canonical gradient satisfies the

condition

$$\begin{aligned} & E\{Ys_2(\epsilon)\} + E\{Y\xi(X,Z,\epsilon)\} \qquad (A.14) \\ = \; & E[\delta\{s_{2*}(\epsilon) + \xi_*(X,Z,\epsilon)\}\{s_2(\epsilon) + \xi(X,Z,\epsilon)\}]. \end{aligned}$$

What is left is to show $E\{Ys_2(\epsilon)\} + E\{Y\xi(X,Z,\epsilon)\} = E\{\phi(X,Z)\}$. This can be shown by

$$\begin{aligned} & E\{Ys_2(\epsilon)\} + E\{Y\xi(X,Z,\epsilon)\} \\ = \; & E\{Ys_2(\epsilon)\} + E\Big\{Y\Big(\big[\phi(X,Z) - E\{\phi(X,Z)|\delta=1\}\big]l(\epsilon) + \\ & E\{\phi(X,Z)|\delta=1\}\frac{\epsilon}{\sigma^2}\Big)\Big\} \\ = \; & E\{Ys_2(\epsilon)\} + E\{Y\phi(X,Z)l(\epsilon)\} - E\{Yl(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\ & + \frac{1}{\sigma^2}E\{\phi(X,Z)|\delta=1\}E(Y\epsilon). \end{aligned}$$

Now using the fact that $Y = r_\vartheta(X) + \gamma(Z) + \epsilon$,

$$\begin{aligned} & E\{Ys_2(\epsilon)\} + E\{Y\phi(X,Z)l(\epsilon)\} - E\{Yl(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\ & + \frac{1}{\sigma^2}E\{\phi(X,Z)|\delta=1\}E(Y\epsilon) \\ = \; & E[\{r_\vartheta(X) + \gamma(Z) + \epsilon\}s_2(\epsilon)] + E[\{r_\vartheta(X) + \gamma(Z) + \epsilon\}\phi(X,Z)l(\epsilon)] \\ & - E[\{r_\vartheta(X) + \gamma(Z) + \epsilon\}l(\epsilon)]E\{\phi(X,Z)|\delta=1\} \\ & + \frac{1}{\sigma^2}E\{\phi(X,Z)|\delta=1\}E[\{r_\vartheta(X) + \gamma(Z) + \epsilon\}\epsilon] \\ = \; & E[\{r_\vartheta(X) + \gamma(Z)\}s_2(\epsilon)] + E\{\epsilon s_2(\epsilon)\} \\ & + E[\{r_\vartheta(X) + \gamma(Z)\}\phi(X,Z)l(\epsilon)] + E\{\epsilon\phi(X,Z)l(\epsilon)\} \\ & - E[\{r_\vartheta(X) + \gamma(Z)\}l(\epsilon)]E\{\phi(X,Z)|\delta=1\} - E\{\epsilon l(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\ & + \frac{1}{\sigma^2}E\{\phi(X,Z)|\delta=1\}E[\{r_\vartheta(X) + \gamma(Z)\}\epsilon] + \frac{1}{\sigma^2}E(\epsilon^2)E\{\phi(X,Z)|\delta=1\}. \end{aligned}$$

Note that since ϵ is independant of X and Z, this equation can be simplified using

Equation A.10 so the terms with $E\{s_2(\epsilon)\} = 0$ and Equation A.11 so the terms with $E\{\epsilon s_2(\epsilon)\} = 0$. In summary we have

$$\begin{aligned} & E\{Y s_2(\epsilon)\} + E\{Y\xi(X,Z,\epsilon)\} \\ = \; & E[\{r_\vartheta(X) + \gamma(Z)\}\phi(X,Z)l(\epsilon)] + E\{\epsilon\phi(X,Z)l(\epsilon)\} \\ & -E[\{r_\vartheta(X) + \gamma(Z)\}l(\epsilon)]E\{\phi(X,Z)|\delta = 1\} - E\{\epsilon l(\epsilon)\}E\{\phi(X,Z)|\delta = 1\} \\ & +\frac{1}{\sigma^2}E\{\phi(X,Z)|\delta = 1\}E[\{r_\vartheta(X) + \gamma(Z)\}\epsilon] + \frac{1}{\sigma^2}E(\epsilon^2)E\{\phi(X,Z)|\delta = 1\}. \end{aligned}$$

Now using the fact that ϵ is independant of X and Z and the fact that $E(\epsilon) = E\{l(\epsilon)\} = 0$ and Lemma VI.7 which states $E\{\epsilon l(\epsilon)\} = 1$ we can simplify the rest of the equation to be

$$\begin{aligned} E\{Y s_2(\epsilon)\} + E\{Y\xi(X,Z,\epsilon)\} &= E\{\phi(X,Z)\} - E\{\phi(X,Z)|\delta = 1\} \\ &\quad +E\{\phi(X,Z)|\delta = 1\} \\ &= E\{\phi(X,Z)\}. \end{aligned}$$

This concludes the simplification of the first half of the equation, so the final form is

$$E\{\phi(X,Z)\} = E[\delta\{s_{2*}(\epsilon) + \xi_*(X,Z,\epsilon)\}\{s_2(\epsilon) + \xi(X,Z,\epsilon)\}]. \qquad \blacksquare$$

Lemma VI.11 *Using the notation given,* $s_{2*}(\epsilon) = 0$ *which means*

$$v_*(Y,X,Z) = \xi_*(X,Z,\epsilon)$$

PROOF: Lemma VI.10 holds for any t and $g(Z)$, so it holds for $t = 0$ and $g(Z) = 0$.

In this case $\phi(X,Z)=0$, and

$$\xi(X,Z,\epsilon)=\Big[\phi(X,Z)-E\{\phi(X,Z)|\delta=1\}\Big]l(\epsilon)+E\{\phi(X,Z)|\delta=1\}\frac{\epsilon}{\sigma^2}=0$$

So Lemma VI.10 in this case means

$$\begin{aligned}
0 &= E[\delta\{s_{2*}(\epsilon)+\xi_*(X,Z,\epsilon)\}s_2(\epsilon)] \\
&= E\{\delta s_{2*}(\epsilon)s_2(\epsilon)+\delta s_2(\epsilon)\xi_*(X,Z,\epsilon)\} \\
&= E\{\delta s_{2*}(\epsilon)s_2(\epsilon)\} \\
&\quad +E\bigg(\delta s_2(\epsilon)\Big[\phi_*(X,Z)l(\epsilon)-E\{\phi_*(X,Z)|\delta=1\}l(\epsilon)+\frac{\epsilon}{\sigma^2}E\{\phi_*(X,Z)|\delta=1\}\Big]\bigg)
\end{aligned}$$

This is simplified in a similar manner to Lemma VI.6 where the expected value of a function of (X,Z,δ) times ϵ is zero.

$$0 = E\{\delta s_{2*}(\epsilon)s_2(\epsilon)\}+E\{\delta s_2(\epsilon)\phi_*(X,Z)l(\epsilon)\}-E[\delta s_2(\epsilon)E\{\phi_*(X,Z)|\delta=1\}l(\epsilon)].$$

Then using Lemma VI.8

$$\begin{aligned}
0 &= E\{\delta s_{2*}(\epsilon)s_2(\epsilon)\}+E(\delta)E\{s_2(\epsilon)\phi_*(X,Z)l(\epsilon)|\delta=1\} \\
&\quad -E(\delta)E\{s_2(\epsilon)l(\epsilon)|\delta=1\}E\{\phi(X,Z)|\delta=1\} \\
&= E\{\delta s_{2*}(\epsilon)s_2(\epsilon)\}+E(\delta)E\{s_2(\epsilon)l(\epsilon)|\delta=1\}E\{\phi(X,Z)|\delta=1\} \\
&\quad -E(\delta)E\{s_2(\epsilon)l(\epsilon)|\delta=1\}E\{\phi(X,Z)|\delta=1\} \\
&= E\{\delta s_{2*}(\epsilon)s_2(\epsilon)\}.
\end{aligned}$$

The solution to this is $s_{2*}(\epsilon)=0$, and $v_*(Y,X,Z)=\xi_*(X,Z,\epsilon)$. ∎

Corollary VI.12 *Lemma VI.10 can be simplified to*

$$E\{\phi(X,Z)\}=E\{\delta\xi_*(X,Z,\epsilon)\xi(X,Z,\epsilon)\}.$$

PROOF: From Lemma VI.11 the equation given by Lemma VI.10 becomes

$$\begin{aligned} E\{\phi(X,Z)\} &= E\left[\delta\xi_*(X,Z,\epsilon)\left\{s_2(\epsilon)+\xi(X,Z,\epsilon)\right\}\right] \\ &= E\{\delta\xi_*(X,Z,\epsilon)s_2(\epsilon)\}+E\{\delta\xi_*(X,Z,\epsilon)\xi(X,Z,\epsilon)\}. \end{aligned}$$

To get the form desired it is sufficient to show $E\{\delta\xi_*(X,Z,\epsilon)s_2(\epsilon)\}=0$.

$$\begin{aligned} & E\{\delta\xi_*(X,Z,\epsilon)s_2(\epsilon)\} \\ = \; & E\left(\delta s_2(\epsilon)\left[\phi_*(X,Z)l(\epsilon)-E\{\phi_*(X,Z)|\delta=1\}l(\epsilon)+\frac{\epsilon}{\sigma^2}E\{\phi_*(X,Z)|\delta=1\}\right]\right) \\ = \; & E\{\delta s_2(\epsilon)\phi_*(X,Z)l(\epsilon)\}-E\{\phi_*(X,Z)|\delta=1\}E\{\delta s_2(\epsilon)l(\epsilon)\} \\ & +\frac{1}{\sigma^2}E\{\phi_*(X,Z)|\delta=1\}E\{\delta s_2(\epsilon)\epsilon\}. \end{aligned}$$

The last term is zero by Equation A.11, the second term can be simplified by Lemma VI.8.

$$\begin{aligned} E\{\delta\xi_*(X,Z,\epsilon)s_2(\epsilon)\} &= E\{\delta\phi_*(X,Z)\}E\{s_2(\epsilon)l(\epsilon)\}-\frac{E(\delta)}{E(\delta)}E\{\delta\phi_*(X,Z)\}E\{s_2(\epsilon)l(\epsilon)\} \\ &= 0. \end{aligned}$$

This finishes the proof. ∎

5. Solving for the canonical gradient

The next Lemma will help simplify the canonical gradient further.

Lemma VI.13 *For the model described above* $E\{\delta\phi_*(X,Z)\}=\sigma^2$.

PROOF: Starting with Corollary VI.12,

$$\begin{aligned}
E\{\phi(X,Z)\} &= E\{\delta\xi_*(X,Z,\epsilon)\xi(X,Z,\epsilon)\} \\
&= E\Bigg(\delta\xi_*(X,Z,\epsilon)\Bigg[\phi(X,Z)l(\epsilon) - E\{\phi(X,Z)|\delta=1\}l(\epsilon) \\
&\quad +\frac{\epsilon}{\sigma^2}E\{\phi(X,Z)|\delta=1\}\Bigg]\Bigg) \\
&= E\{\delta\xi_*(X,Z,\epsilon)\phi(X,Z)l(\epsilon)\} - E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\}E\{\phi(X,Z)|\delta=1\} \\
&\quad +\frac{1}{\sigma^2}E\{\delta\xi_*(X,Z,\epsilon)\epsilon\}E\{\phi(X,Z)|\delta=1\}.
\end{aligned}$$

By Lemma VI.8

$$\begin{aligned}
E\{\phi(X,Z)\} &= E\{\delta\xi_*(X,Z,\epsilon)\phi(X,Z)l(\epsilon)\} - E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\}\frac{E\{\delta\phi(X,Z)\}}{E(\delta)} \\
&\quad +\frac{1}{\sigma^2}E\{\delta\xi_*(X,Z,\epsilon)\epsilon\}\frac{E\{\delta\phi(X,Z)\}}{E(\delta)} \\
&= E\Bigg(\Bigg[\delta\xi_*(X,Z,\epsilon)l(\epsilon) - \delta\frac{E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\}}{E(\delta)} \\
&\quad +\delta\frac{E\{\delta\xi_*(X,Z,\epsilon)\epsilon\}}{\sigma^2E(\delta)}\Bigg]\phi(X,Z)\Bigg).
\end{aligned}$$

By setting the equation to zero the form is

$$\begin{aligned}
0 &= E\Bigg(\Bigg[\delta\xi_*(X,Z,\epsilon)l(\epsilon) - \delta\frac{E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\}}{E(\delta)} \\
&\quad +\delta\frac{E\{\delta\xi_*(X,Z,\epsilon)\epsilon\}}{\sigma^2E(\delta)} - 1\Bigg]\phi(X,Z)\Bigg). \qquad \text{(A.15)}
\end{aligned}$$

Since this must be true for any $\phi(X,Z)$ including $\phi(X,Z)=1$,

$$\begin{aligned}
0 &= E\Bigg[\delta\xi_*(X,Z,\epsilon)l(\epsilon) - \delta\frac{E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\}}{E(\delta)} + \delta\frac{E\{\delta\xi_*(X,Z,\epsilon)\epsilon\}}{\sigma^2E(\delta)} - 1\Bigg] \\
&= E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\} - \frac{E(\delta)}{E(\delta)}E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\} + \frac{E(\delta)}{\sigma^2E(\delta)}E\{\delta\xi_*(X,Z,\epsilon)\epsilon\} - 1 \\
&= \frac{1}{\sigma^2}E\{\delta\xi_*(X,Z,\epsilon)\epsilon\} - 1.
\end{aligned}$$

Which means

$$\begin{aligned} 1 &= \frac{1}{\sigma^2}E\left(\delta\epsilon\left[\phi_*(X,Z)l(\epsilon) - E\{\phi_*(X,Z)|\delta=1\}l(\epsilon) + \frac{\epsilon}{\sigma^2}E\{\phi_*(X,Z)|\delta=1\}\right]\right) \\ &= \frac{1}{\sigma^2}E\{\delta\epsilon\phi_*(X,Z)l(\epsilon)\} - \frac{1}{\sigma^2}E\{\delta\epsilon l(\epsilon)\}E\{\phi_*(X,Z)|\delta=1\} \\ &\quad + \frac{1}{\sigma^2}E\left\{\delta\frac{\epsilon^2}{\sigma^2}\right\}E\{\phi_*(X,Z)|\delta=1\}. \end{aligned}$$

By Lemma VI.7 and using Lemma VI.8,

$$\begin{aligned} \sigma^2 &= E\{\delta\phi_*(X,Z)\} - E(\delta)E\{\phi_*(X,Z)|\delta=1\} + E(\delta)E\{\phi_*(X,Z)|\delta=1\} \\ &= E\{\delta\phi_*(X,Z)\}. \end{aligned}$$

■

One benefit from this lemma is in the simplification of $\xi_*(X,Z,\epsilon)$ as shown in the following Corollary.

Corollary VI.14 *Using the notation given*

$$\xi_*(X,Z,\epsilon) = \phi_*(X,Z)l(\epsilon) - \frac{\sigma^2}{E(\delta)}l(\epsilon) + \frac{\epsilon}{E(\delta)}.$$

Proof: From Lemma VI.8 substituting in Lemma VI.13.

$$E\{\phi_*(X,Z)|\delta=1\} = \frac{\sigma^2}{E(\delta)}.$$

Which means

$$\begin{aligned} \xi_*(X,Z,\epsilon) &= \phi_*(X,Z)l(\epsilon) - E\{\phi_*(X,Z)|\delta=1\}l(\epsilon) - \frac{\epsilon}{\sigma^2}E\{\phi_*(X,Z)|\delta=1\} \\ &= \phi_*(X,Z)l(\epsilon) - \frac{\sigma^2}{E(\delta)}l(\epsilon) + \frac{\epsilon}{E(\delta)}. \end{aligned}$$

■

The simplification of Equation A.15 using Lemma VI.13 will be easier to follow if the expected values are solved individually. This will be done in the following three lemmas. Define $\mathbb{I}$ as the Fischer Information for ϵ.

Lemma VI.15 *Using the notation above*

$$E\{\delta l(\epsilon)\xi_*(X,Z,\epsilon)\} = 1.$$

PROOF:

$$\begin{aligned} E\{\delta l(\epsilon)\xi_*(X,Z,\epsilon)\} &= E\left[\delta l(\epsilon)\left\{\phi_*(X,Z)l(\epsilon) - \frac{\sigma^2}{E(\delta)}l(\epsilon) + \frac{\epsilon}{E(\delta)}\right\}\right] \\ &= E\{\delta\phi_*(X,Z)\}E\{l^2(\epsilon)\} - \frac{\sigma^2}{E(\delta)}E(\delta)E\{l^2(\epsilon)\} + \frac{E(\delta)}{E(\delta)}E\{\epsilon l(\epsilon)\} \\ &= E\{\delta\phi_*(X,Z)\}\mathbb{I} - \frac{E(\delta)}{E(\delta)}\sigma^2\mathbb{I} + 1. \end{aligned}$$

By Lemma VI.13

$$\begin{aligned} E\{\delta l(\epsilon)\xi_*(X,Z,\epsilon)\} &= \sigma^2\mathbb{I} - \sigma^2\mathbb{I} + 1 \\ &= 1. \end{aligned}$$

■

Lemma VI.16 *Using the notation above*

$$E\{\delta\epsilon\xi_*(X,Z,\epsilon)\} = \sigma^2.$$

PROOF:

$$\begin{aligned} E\{\delta\epsilon\xi_*(X,Z,\epsilon)\} &= E\left[\delta\epsilon\left\{\phi_*(X,Z)l(\epsilon) - \frac{\sigma^2}{E(\delta)}l(\epsilon) + \frac{\epsilon}{E(\delta)}\right\}\right] \\ &= E\{\delta\phi_*(X,Z)\}E\{\epsilon l(\epsilon)\} - \frac{\sigma^2}{E(\delta)}E(\delta)E\{\epsilon l(\epsilon)\} + \frac{E(\delta)}{E(\delta)}E(\epsilon^2). \end{aligned}$$

By Lemma VI.7 and Lemma VI.13

$$\begin{aligned} E\{\delta\epsilon\xi_*(X,Z,\epsilon)\} &= E\{\delta\phi_*(X,Z)\} - \sigma^2 + \sigma^2 \\ &= \sigma^2. \end{aligned}$$

■

Lemma VI.17 *Using the notation above*

$$\begin{aligned} E\{\delta\phi(X,Z)l(\epsilon)\xi_*(X,Z,\epsilon)\} &= E\{\delta\phi(X,Z)\phi_*(X,Z)\}\mathbb{I} - \frac{\sigma^2}{E(\delta)}E\{\delta\phi(X,Z)\}\mathbb{I} \\ &\quad + \frac{E\{\delta\phi(X,Z)\}}{E(\delta)}. \end{aligned}$$

PROOF:

$$\begin{aligned} & E\{\delta\phi(X,Z)l(\epsilon)\xi_*(X,Z,\epsilon)\} \\ = \; & E\left[\delta\phi(X,Z)l(\epsilon)\left\{\phi_*(X,Z)l(\epsilon) - \frac{\sigma^2}{E(\delta)}l(\epsilon) + \frac{\epsilon}{E[\delta]}\right\}\right] \\ = \; & E\{\delta\phi(X,Z)\phi_*(X,Z)\}E\{l^2(\epsilon)\} - \frac{\sigma^2}{E(\delta)}E\{\delta\phi(X,Z)\}E\{l^2(\epsilon)\} \\ & + \frac{E\{\delta\phi(X,Z)\}}{E(\delta)}E\{\epsilon l(\epsilon)\} \\ = \; & E\{\delta\phi(X,Z)\phi_*(X,Z)\}\mathbb{I} - \frac{\sigma^2}{E(\delta)}E\{\delta\phi(X,Z)\}\mathbb{I} + \frac{E\{\delta\phi(X,Z)\}}{E(\delta)}. \end{aligned}$$

■

Now an important lemma to solve for $g_*(Z)$ in terms of t_*.

Lemma VI.18 *Using the notation given*

$$g_*(Z) = \frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} - {t_*}^\top \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}.$$

PROOF: Begin with Equation A.15

$$\begin{aligned} 0 &= E\Bigg(\Bigg[\delta\xi_*(X,Z,\epsilon)l(\epsilon) - \delta\frac{E\{\delta\xi_*(X,Z,\epsilon)l(\epsilon)\}}{E(\delta)} \\ &\quad +\delta\frac{E\{\delta\xi_*(X,Z,\epsilon)\epsilon\}}{\sigma^2 E(\delta)} - 1\Bigg]\phi(X,Z)\Bigg). \end{aligned}$$

Using Lemma VI.15 and Lemma VI.16,

$$\begin{aligned} 0 &= E\left[\left\{\delta\xi_*(X,Z,\epsilon)l(\epsilon) - \delta\frac{1}{E(\delta)} + \delta\frac{\sigma^2}{\sigma^2 E(\delta)} - 1\right\}\phi(X,Z)\right] \\ &= E\left[\left\{\delta\xi_*(X,Z,\epsilon)l(\epsilon) - 1\right\}\phi(X,Z)\right]. \end{aligned}$$

Then using Lemma VI.17,

$$\begin{aligned} 0 &= E\{\delta\phi(X,Z)\phi_*(X,Z)\}\mathbb{I} - \frac{\sigma^2}{E(\delta)}E\{\delta\phi(X,Z)\}\mathbb{I} + \frac{E\{\delta\phi(X,Z)\}}{E(\delta)} - E\{\phi(X,Z)\} \\ &= E\left[\left\{\delta\phi_*(X,Z)\mathbb{I} - \frac{\sigma^2\delta\mathbb{I}}{E(\delta)} + \frac{\delta}{E(\delta)} - 1\right\}\phi(X,Z)\right]. \end{aligned} \tag{A.16}$$

This is true for any t, so set $t = 0$. In this case $\phi(X,Z) = g(Z)$. The equation then becomes

$$0 = E\left[\left\{\delta\phi_*(X,Z)\mathbb{I} - \frac{\sigma^2\delta\mathbb{I}}{E(\delta)} + \frac{\delta}{E(\delta)} - 1\right\}g(Z)\right].$$

Using the Law of Iterated Expectations I will rewrite this equation so that $\phi(X,Z)$ is being multiplied by a function of Z.

$$0 = E\left[\left\{E\{\delta\phi_*(X,Z)|Z\}\mathbb{I} - \frac{\sigma^2 E(\delta|Z)\mathbb{I}}{E(\delta)} + \frac{E(\delta|Z)}{E(\delta)} - 1\right\}g(Z)\right]. \tag{A.17}$$

To simplify the notation momentarily let

$$M(Z) = E\{\delta\phi_*(X,Z)|Z\}\mathbb{I} - \frac{\sigma^2 E(\delta|Z)\mathbb{I}}{E(\delta)} + \frac{E(\delta|Z)}{E(\delta)} - 1$$

so that Equation A.17 becomes

$$0 = E\{M(Z)g(Z)\}. \tag{A.18}$$

Note also that

$$\begin{aligned} E[M(Z)] &= E\left[E\{\delta\phi_*(X,Z)|Z\}\mathbb{I} - \frac{\sigma^2 E(\delta|Z)\mathbb{I}}{E(\delta)} + \frac{E(\delta|Z)}{E(\delta)} - 1\right] \\ &= E\{\delta\phi_*(X,Z)\}\mathbb{I} - \frac{\sigma^2 E(\delta)\mathbb{I}}{E(\delta)} + \frac{E(\delta)}{E(\delta)} - 1 \\ &= E\{\delta\phi_*(X,Z)\}\mathbb{I} - \sigma^2\mathbb{I} \\ &= 0. \end{aligned}$$

This is true for any $g(Z)$, so let $g(Z) = E\{g(Z)\} + g_0(Z)$ where $E\{g(Z)\}$ is a constant, and $E\{g_0(Z)\} = 0$. This changes Equation A.18 to

$$\begin{aligned} 0 &= E\Big(M(Z)[E\{g(Z)\} + g_0(Z)]\Big) \\ &= E[M(Z)E\{g(Z)\}] + E\{M(Z)g_0(Z)\} \\ &= E\{g(Z)\}E\{M(Z)\} + E\{M(Z)g_0(Z)\} \\ &= E\{M(Z)g_0(Z)\}. \end{aligned}$$

This is true for all $g_0(Z)$, which is not a constant, then $M(Z)$ must equal zero.

$$\begin{aligned} 0 &= M(Z) \\ &= E\{\delta\phi_*(X,Z)|Z\}\mathbb{I} - \frac{\sigma^2 E(\delta|Z)\mathbb{I}}{E(\delta)} + \frac{E(\delta|Z)}{E(\delta)} - 1. \end{aligned}$$

Then substituting $\phi_*(X,Z) = {t_*}^\top \dot{r}_\vartheta(X) + g_*(Z)$

$$\begin{aligned} 0 &= E[\delta\{{t_*}^\top \dot{r}_\vartheta(X) + g_*(Z)\}|Z]\mathbb{I} - \frac{\sigma^2 E(\delta|Z)\mathbb{I}}{E(\delta)} + \frac{E(\delta|Z)}{E(\delta)} - 1 \\ &= {t_*}^\top E\{\delta\dot{r}_\vartheta(X)|Z\}\mathbb{I} + g_*(Z)E(\delta|Z)\mathbb{I} - \frac{\sigma^2 E(\delta|Z)\mathbb{I}}{E(\delta)} + \frac{E(\delta|Z)}{E(\delta)} - 1. \end{aligned}$$

Which means

$$g_*(Z) = \frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} - {t_*}^\top \frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}.$$ ∎

Now the solution for t_* can be found.

Lemma VI.19 *Using the given notation*

$$\begin{aligned} t_* &= \frac{1}{\mathbb{I}}\left(E\{\delta\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta\dot{r}_\vartheta(X)|Z\}E\{\delta\dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right]\right)^{-1} \\ &\quad \left[E\{\dot{r}_\vartheta(X)\} - \frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right]. \end{aligned} \tag{A.19}$$

Proof: Begin with Equation A.16,

$$0 = E\left[\left\{\delta\phi_*(X,Z)\mathbb{I} - \frac{\sigma^2\delta\mathbb{I}}{E(\delta)} + \frac{\delta}{E(\delta)} - 1\right\}\phi(X,Z)\right]$$

which is true for any $g(Z)$, so set $g(Z) = 0$, which means $\phi(X,Z) = t^\top\dot{r}_\vartheta(X)$, and

$$\begin{aligned} 0 &= E\left[\left\{\delta\phi_*(X,Z)\mathbb{I} - \frac{\sigma^2\delta\mathbb{I}}{E(\delta)} + \frac{\delta}{E(\delta)} - 1\right\}t^\top\dot{r}_\vartheta(X)\right] \\ &= E\left\{\delta\phi_*(X,Z)\mathbb{I}t^\top\dot{r}_\vartheta(X) - \frac{\sigma^2\delta\mathbb{I}t^\top\dot{r}_\vartheta(X)}{E(\delta)} + \frac{\delta t^\top\dot{r}_\vartheta(X)}{E(\delta)} - t^\top\dot{r}_\vartheta(X)\right\}. \end{aligned}$$

Then substituting $\phi_*(X,Z) = {t_*}^\top \dot{r}_\vartheta(X) + g_*(Z)$,

$$\begin{aligned} 0 &= E\Bigg[\delta\{{t_*}^\top \dot{r}_\vartheta(X) + g_*(Z)\}\mathbb{I}t^\top \dot{r}_\vartheta(X) \\ &\quad -\frac{\sigma^2\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)}{E(\delta)} + \frac{\delta t^\top \dot{r}_\vartheta(X)}{E(\delta)} - t^\top \dot{r}_\vartheta(X)\Bigg] \\ &= E\Bigg\{\delta {t_*}^\top \dot{r}_\vartheta(X)\mathbb{I}t^\top \dot{r}_\vartheta(X) + \delta g_*(Z)\mathbb{I}t^\top \dot{r}_\vartheta(X) - \frac{\sigma^2\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)}{E(\delta)} \\ &\quad +\frac{\delta t^\top \dot{r}_\vartheta(X)}{E(\delta)} - t^\top \dot{r}_\vartheta(X)\Bigg\}. \end{aligned}$$

Using Lemma VI.18,

$$\begin{aligned} 0 &= E\Bigg(\delta {t_*}^\top \dot{r}_\vartheta(X)\mathbb{I}t^\top \dot{r}_\vartheta(X) \\ &\quad +\delta\left[\frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} - {t_*}^\top \frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right]\mathbb{I}t^\top \dot{r}_\vartheta(X) \\ &\quad -\frac{\sigma^2\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)}{E(\delta)} + \frac{\delta t^\top \dot{r}_\vartheta(X)}{E(\delta)} - t^\top \dot{r}_\vartheta(X)\Bigg) \\ &= E\Bigg[\delta {t_*}^\top \dot{r}_\vartheta(X)\mathbb{I}t^\top \dot{r}_\vartheta(X) + \frac{\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)\sigma^2}{E(\delta)} - \frac{\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)}{E(\delta)\mathbb{I}} + \frac{\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)}{E(\delta|Z)\mathbb{I}} \\ &\quad -\delta\mathbb{I}t^\top \dot{r}_\vartheta(X){t_*}^\top \frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} - \frac{\sigma^2\delta\mathbb{I}t^\top \dot{r}_\vartheta(X)}{E(\delta)} + \frac{\delta t^\top \dot{r}_\vartheta(X)}{E(\delta)} - t^\top \dot{r}_\vartheta(X)\Bigg] \\ &= \mathbb{I}t^\top E\{\delta\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\}t_* + \sigma^2\mathbb{I}t^\top \frac{E\{\delta\dot{r}_\vartheta(X)\}}{E(\delta)} \\ &\quad -t^\top \frac{E\{\delta\dot{r}_\vartheta(X)\}}{E(\delta)} + t^\top E\left\{\frac{\delta\dot{r}_\vartheta(X)}{E(\delta|Z)}\right\} \\ &\quad -\mathbb{I}t^\top E\left[\delta\dot{r}_\vartheta(X)\frac{E\{\delta\dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right]t_* - \sigma^2\mathbb{I}t^\top \frac{E\{\delta\dot{r}_\vartheta(X)\}}{E(\delta)} + t^\top \frac{E\{\delta\dot{r}_\vartheta(X)\}}{E(\delta)} \\ &\quad -t^\top E\{\dot{r}_\vartheta(X)\}. \end{aligned}$$

Then factoring out the constants t and t_*,

$$\begin{aligned}
0 &= t^\top \Bigg\{ \sigma^2 \mathbb{I} \frac{E\{\delta \dot{r}_\vartheta(X)\}}{E(\delta)} - \frac{E\{\delta \dot{r}_\vartheta(X)\}}{E(\delta)} + E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right] - \sigma^2 \mathbb{I} \frac{E\{\delta \dot{r}_\vartheta(X)\}}{E(\delta)} \\
&\quad + \frac{E\{\delta \dot{r}_\vartheta(X)\}}{E(\delta)} - E\{\dot{r}_\vartheta(X)\} \\
&\quad + \left(\mathbb{I} E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - \mathbb{I} E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E\{\delta|Z\}}\right] \right) t_* \Bigg\} \\
&= t^\top \Bigg\{ E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right] - E\{\dot{r}_\vartheta(X)\} \\
&\quad + \left(\mathbb{I} E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - \mathbb{I} E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right] \right) t_* \Bigg\}.
\end{aligned}$$

Since this is true for any t let $t = 1$. Then

$$\begin{aligned}
0 &= E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right] - E\{\dot{r}_\vartheta(X)\} \\
&\quad + \left(\mathbb{I} E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - \mathbb{I} E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right] \right) t_*.
\end{aligned}$$

This means

$$\begin{aligned}
t_* &= \frac{1}{\mathbb{I}} \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right] \right)^{-1} \\
&\quad \left(E\{\dot{r}_\vartheta(X)\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right] \right).
\end{aligned}$$

■

Lemma VI.20 *Using the notation given*

$$\begin{aligned}
v_*(Y, X, Z) &= \frac{\epsilon}{E(\delta)} - \frac{l(\epsilon)}{E(\delta)\mathbb{I}} + \frac{l(\epsilon)}{E(\delta|Z)\mathbb{I}} - \\
&\quad \frac{l(\epsilon)}{\mathbb{I}} \left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right] \right) \\
&\quad \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right] \right)^{-1} \\
&\quad \left[\dot{r}_\vartheta(X) - \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right].
\end{aligned}$$

PROOF: The proof requires plugging t and $g(Z)$ into Lemma VI.11. Using Lemma VI.19 with Lemma VI.18

$$
\begin{aligned}
g_*(Z) &= \frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} \\
&\quad -\left\{ \frac{1}{\mathbb{I}} \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)} \right] \right)^{-1} \right. \\
&\quad \left. \left(E\{\dot{r}_\vartheta(X)\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right] \right) \right\}^\top \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \\
&= \frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} \\
&\quad -\frac{1}{\mathbb{I}} \left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right] \right) \\
&\quad \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)} \right] \right)^{-1} \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}.
\end{aligned}
$$

Then

$$\begin{aligned}
\phi_*(X,Z) &= t_*{}^\top \dot{r}_\vartheta(X) + g_*(Z) \\
&= \left\{ \frac{1}{\mathbb{I}} \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)} \right] \right)^{-1} \right. \\
&\quad \left. \left(E\{\dot{r}_\vartheta(X)\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right] \right) \right\}^\top \dot{r}_\vartheta(X) \\
&\quad + \frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} \\
&\quad - \frac{1}{\mathbb{I}} \left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right] \right) \\
&\quad \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - \right. \\
&\quad \left. E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)} \right] \right)^{-1} \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \\
&= \frac{\sigma^2}{E(\delta)} - \frac{1}{E(\delta)\mathbb{I}} + \frac{1}{E(\delta|Z)\mathbb{I}} - \frac{1}{\mathbb{I}} \left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right] \right) \\
&\quad \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)} \right] \right)^{-1} \\
&\quad \left[\dot{r}_\vartheta(X) - \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right].
\end{aligned}$$

Then using Corollary VI.14

$$\begin{aligned}
\xi_*(X,Z,\epsilon) &= \frac{\epsilon}{E(\delta)} - \frac{l(\epsilon)}{E(\delta)\mathbb{I}} + \frac{l(\epsilon)}{E(\delta|Z)\mathbb{I}} - \frac{l(\epsilon)}{\mathbb{I}} \left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right] \right) \\
&\quad \left(E\{\delta \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta \dot{r}_\vartheta(X)|Z\} E\{\delta \dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)} \right] \right)^{-1} \\
&\quad \left[\dot{r}_\vartheta(X) - \frac{E\{\delta \dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)} \right].
\end{aligned}$$

The final form is now straightforward by Lemma VI.11. ∎

Theorem VI.21 *For the model*

$$Y = \vartheta^\top r_\vartheta(X) + \gamma(Z) + \epsilon$$

where Y is the response, $r_\vartheta(X)$ is the known parametric function and ϑ is the unknown parameter for the random covariate X with dimension k_1, $\gamma(Z)$ is the unknown non-parametric component for the random covariate Z with dimension k_2, and ϵ is the random error term with mean 0 and variance σ^2. Further assume (X, Z) is independent of ϵ. The responses are MAR with respect to the variable δ such that the observed data is $(X, Z, \delta, \delta Y)$. The canonical gradient for estimating $E(Y)$ is

$$\begin{aligned} gr_{(\vartheta_*,\gamma_*)} &= r_\vartheta(X) + \gamma(Z) - E\{r_\vartheta(X) + \gamma(Z)\} + \frac{\delta\epsilon}{E(\delta)} - \frac{\delta l(\epsilon)}{E(\delta)\mathbb{I}} + \frac{\delta l(\epsilon)}{E(\delta|Z)\mathbb{I}} \\ &\quad -\frac{\delta l(\epsilon)}{\mathbb{I}}\left(E\{\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right]\right) \\ &\quad \left(E\{\delta\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\} - E\left[\frac{E\{\delta\dot{r}_\vartheta(X)|Z\}E\{\delta\dot{r}_\vartheta(X)|Z\}^\top}{E(\delta|Z)}\right]\right)^{-1} \\ &\quad \left[\dot{r}_\vartheta(X) - \frac{E\{\delta\dot{r}_\vartheta(X)|Z\}}{E(\delta|Z)}\right]. \end{aligned}$$

PROOF: The proof follows directly from Lemma VI.4 and Lemma VI.20 ∎

This canonical gradient can be used to check for whether an estimator is asymptotically efficient. If the influence function of the estimator matches this canonical gradient it is efficient. Unfortunately for this example an estimator has yet to be found which achieves this level of efficiency. The complexity can be observed by comparing this formula with Lemma II.2 where the model was a simple linear model with normally distributed errors.

VITA

Scott Crawford was born in Cedar City, Utah. He received his B.S. in mathematics from Southern Utah University and M.S. in statistics from Brigham Young University in May 2004 and May 2006, respectively. He received his Ph.D. in statistics from Texas A&M University in August 2012. He has accepted a position at the University of Wyoming as a lecturer to begin in August 2012. His current research interests lie in efficiency, missing data models, actuarial science, asymptotic theory, and mass spectrometry.

His permanent address is:

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX, 77843-3143. Email: crawford@stat.tamu.edu

The typist for this thesis was Scott Crawford.